2017

Development and evaluation of a model to correct tapered element oscillating microbalance (TEOM) readings of PM2.5 in Chullora, Sydney.

Alexandra E. Northam

Development and evaluation of a model to correct tapered element oscillating microbalance (TEOM) readings of PM2.5 in Chullora, Sydney.

Abstract

Tapered element oscillating microbalance (TEOM) monitors offer substantial benefits to air pollution regulatory bodies in regards to their reduced need for labor and their ability to provide data in “real-time” through an automated system. However, research has demonstrated that the TEOM tends to provide inaccurate particulate matter concentrations due to the operational framework of the instrument. This paper presents the results of collocated comparisons of two $PM_{2.5}$ monitors, a TEOM and a beta-attenuation monitor (BAM), conducted from September 2010 to November 2012, in the greater urban area of Chullora, Sydney, Australia. The objective of this work is to define the relationship between these two monitors, and develop a model to correct the TEOM instrument to bring in into line with what is seen as the ‘gold standard’ of $PM_{2.5}$ monitors, the BAM. The results show that there is a significant positive linear relationship between TEOM and BAM samplers, at an hourly and daily scale (*p-value* < 0.001), with the TEOM generally reporting lower $PM_{2.5}$ concentrations than the collocated BAM. Local meteorological, air pollution and gas covariates were integrated into a single linear model for $PM_{2.5}$ predictions, at hourly and daily intervals. Although the model significantly improved the R^2 of the agreement between instruments at hourly intervals (from 0.24 to 0.43, with a 95% confidence interval of 6.97 μg/m^3 and 7.58 μg/m^3), results indicate autocorrelation in the residuals of the model, suggesting there is information in the residuals that should be included in computing the forecast. Hence, producing a robust hourly model remains a challenge. A model for daily predictions improved the agreement between instruments (R^2 improved from 0.75 to 0.81, with a 95% confidence interval of 7.93 μg/m^3 and 8.52 μg/m^3). Time series cross validation demonstrated a strong statistical performance of the daily model on independent data (*FAC2* = 1.00*, mean bias* = 0.02 μg/m^3*, Pearson’s correlation coefficient* = 0.92). A 7-year record of hourly TEOM measurements from 2004 to 2010 were corrected, based on the equation derived from the daily 2-year collocated measurements. Although not significantly significant, the overall trend analysis combining both the adjusted TEOM and BAM measurements demonstrated 0.62% per year increase (95% confidence interval of -0.53%, 2.03%) in $PM_{2.5}$ concentrations from 2004 to 2012. Only spring produced a statistically significant increase in $PM_{2.5}$ concentrations from 2004 to 2012, of 4.93% per year (3.41%, 6.1%). Hence, our daily model can robustly estimate historical $PM_{2.5}$ concentrations at Chullora when $PM_{2.5}$ BAM measurements were not available.

Implications: The robustness of the daily model means that it can be applied to correct the historical TEOM data, to examine long term-trends at this site. This technique of correction can be adapted to other sites in Sydney, serving as a stepping stone in the long term-goal of developing an Air Quality Index for New South Wales, for periods when a TEOM was the only $PM_{2.5}$ sampler at a site.

Degree Type

Thesis

Degree Name

BEnviSc Hons

Department

School of Earth & Environmental Science

Development and evaluation of a model to correct tapered element oscillating microbalance (TEOM) readings of $PM_{2.5}$ in Chullora, Sydney.

ALEXANDRA ELLIE NORTHAM

A research report submitted in partial fulfillment of the requirements for the award degree of Bachelor of Environmental Science (Honours) in the School of Earth and Environmental Science Faculty of Science, Medicine and Health, The University of Wollongong, 2017.

24th October 2017

The information in this thesis is entirely the result of investigations conducted by the author, unless otherwise acknowledged, and has not been submitted in part, or otherwise, for any other degree or qualification.

Signed: …… **Date:** 24/10/2017

Acknowledgements

I would like to extend a massive thankyou to everyone who has helped throughout the course of this project. I am especially thrilled to have worked with such driven and intelligent women in the science field.

Thankyou to Sandy Burden for your guidance, and the many hours you dedicated to helping me wrap my head around the statistics behind the project. I would not have been able to complete this without you. Thankyou to Clare Murphy for providing useful feedback, and helping to plan and organize this project. I'd also like to thank my external supervisors, Yvonne Scorgie and Lisa Tzu-Chi Chang from the Office of Environment and Heritage, for your ongoing interest throughout the year, and for organizing the logistics behind the project. And thankyou to Elise-Andree Guerette for introducing me to the world of coding. Thankyou for taking the time to show me the basics of R.

I would like to express my sincere appreciation to my family and friends who have made this year possible. James, thankyou for the ongoing love and support throughout the year. And to my mum, Christine and my sister, Cait, for your encouragement and support not only this year, but through my whole degree. Also, I'd like to thank Hussain for the constant motivation and coffee breaks together.

Lastly, thankyou to the University of Wollongong for their rigorous academic program and their established partnerships with significant industry players, providing me with priceless opportunities to further develop my skills and passion for the science field.

I would like to acknowledge and pay respect to the Bidjigal people and the Wadi Wadi people of the Dharwal nation, on whose land this research was performed. I respect the long and continuing relationship between Indigenous people and their country.

Abstract

Tapered element oscillating microbalance (TEOM) monitors offer substantial benefits to air pollution regulatory bodies in regards to their reduced need for labor and their ability to provide data in "real-time" through an automated system. However, research has demonstrated that the TEOM tends to provide inaccurate particulate matter concentrations due to the operational framework of the instrument. This paper presents the results of collocated comparisons of two $PM_{2.5}$ monitors, a TEOM and a beta-attenuation monitor (BAM), conducted from September 2010 to November 2012, in the greater urban area of Chullora, Sydney, Australia. The objective of this work is to define the relationship between these two monitors, and develop a model to correct the TEOM instrument to bring in into line with what is seen as the 'gold standard' of $PM_{2.5}$ monitors, the BAM. The results show that there is a significant positive linear relationship between TEOM and BAM samplers, at an hourly and daily scale (*p-value* < 0.001), with the TEOM generally reporting lower $PM_{2.5}$ concentrations than the collocated BAM. Local meteorological, air pollution and gas covariates were integrated into a single linear model for $PM_{2.5}$ predictions, at hourly and daily intervals. Although the model significantly improved the R^2 of the agreement between instruments at hourly intervals (from 0.24 to 0.43, with a 95% confidence interval of 6.97 $\mu g/m^3$ and 7.58 $\mu g/m^3$), results indicate autocorrelation in the residuals of the model, suggesting there is information in the residuals that should be included in computing the forecast. Hence, producing a robust hourly model remains a challenge. A model for daily predictions improved the agreement between instruments (R^2 improved from 0.75 to 0.81, with a 95% confidence interval of 7.93 $\mu g/m^3$ and 8.52 $\mu g/m^3$). Time series cross validation demonstrated a strong statistical performance of the daily model on independent data (*FAC2* = 1.00*, mean bias* = 0.02 $\mu g/m^3$*, Pearson's correlation coefficient* = 0.92). A 7-year record of hourly TEOM measurements from 2004 to 2010 were corrected, based on the equation derived from the daily 2-year collocated measurements. Although not significantly significant, the overall trend analysis combining both the adjusted TEOM and BAM measurements demonstrated 0.62% per year increase (95% confidence interval of -0.53%, 2.03%) in $PM_{2.5}$ concentrations from 2004 to 2012. Only spring produced a statistically significant increase in $PM_{2.5}$ concentrations from 2004 to 2012, of 4.93% per year (3.41%, 6.1%). Hence, our daily model can robustly estimate historical $PM_{2.5}$ concentrations at Chullora when $PM_{2.5}$ BAM measurements were not available.

Implications: The robustness of the daily model means that it can be applied to correct the historical TEOM data, to examine long term-trends at this site. This technique of correction can be adapted to other sites in Sydney, serving as a stepping stone in the long term-goal of developing an Air Quality Index for New South Wales, for periods when a TEOM was the only $PM_{2.5}$ sampler at a site.

Table of contents

List of figures

List of tables

Chapter 1: Introduction

1.1 Overview

One important pollutant that affects air quality in urban and rural areas is particulate matter (PM) (or aerosols). To understand the full effects of PM on health and the climate system, routine monitoring of PM is necessary. There are a range of instruments available to measure PM, some measuring more accurately than others. A tapered element oscillating microbalance (TEOM 1400AB), which measures $PM_{2.5}$, was operational at the Chullora air quality monitoring site from 2004 to 2012. This instrument is known to underestimate the measurement of PM due the heating mechanism causing a loss of semi-volatile material and ammonium nitrate from the fine PM fraction, resulting in an incorrect measurement being recorded (Ayers et al., 1999) A beta-attenuation monitor (BAM 5014i) was collocated with the TEOM, between 2010 and 2012. Since BAMs are seen as the 'gold standard' for recording $PM_{2.5}$, the data recorded over the collocated period (including meteorological and other pollutant data) is used to develop a model to 'correct' the TEOM values, bringing them into line with the BAM, prior to the BAMs being in place. A variety of papers and organisations have developed methods to account for the underestimation of $PM_{2.5}$ mass concentration, but none have been calculated for the Sydney region.

Understanding the comparability of data from the TEOM and BAM when operating in the field is of prime importance. It will enable a long-term consistent record of $PM_{2.5}$ to be established at Chullora, dating from 2004 to 2017. It will provide us with an indication of how the distribution of $PM_{2.5}$ has changed over time. As adequately summarized by Blanchard et al. (2011, p.339);

> *"A long record with greater spatial coverage is of value both for detecting trends and for assessing temporal and spatial variations in exposures to air pollutants, a crucial step in developing a quantitative understanding of the effects of specific pollutants on particular health endpoints."*

1.2 Characterisation of Particulate Matter

The study of urban air pollution involves monitoring a suite of variables, one of the most important being atmospheric PM. PM is the total of all solid and liquid particles present in the atmosphere. The chemical and physical characteristics of PM are complex, and their size, shape, physical, chemical and thermodynamic properties can vary according to local

sources, source strength and atmospheric processes (Kam, 2012). PM can appear in the atmosphere as a result of photo-chemically formed particles, mechanical processes, wind erosion, bushfires, gasoline and diesel combustion, biogenic emissions and sea salt. The detrimental effects PM has on human health, visibility and the climate is well established in the literature (World Health Organization, 2016, Morgan et al., 2013, Wark et al., 1998). Hence, an accurate reading of PM is necessary to infer the potential extent of these effects, for policy makers and the community.

Particles can be present in ambient air as an outcome of primary or secondary processes. Particles directly emitted into the atmosphere are termed primary pollutants (i.e. fugitive dust or ash), whereas particles formed through chemical reactions of other pollutants in the atmosphere are termed secondary pollutants (i.e. photochemical reactions with combustion gasses). PM is mainly produced through secondary processes (Australian Government Department of Environment and Energy, 2014). Ambient PM is not one specific pollutant, but consists of a number of chemical species, including elemental carbon, inorganic ions (nitrate and sulphate), trace metals (toxic, crustal and transition metals), and a range of organic species.

The aerodynamic diameter of PM influences its behaviour. PM of 10 microns (μm) in aerodynamic diameter or less is recognised as PM_{10}. PM of 2.5 microns or less is recognised as $PM_{2.5}$. PM_{10} and $PM_{2.5}$ are used to define coarse and fine particulate matter. PM between PM_{10} and $PM_{2.5}$ describes coarse PM, and $PM_{2.5}$ or less describes fine PM (Sienfeld and Pandis, 1998). Particulates larger than 10 microns are of less concern than those less than 10 microns, as they have a lower residency time in the atmosphere and are less likely to have a detrimental effect on human health. Smaller particles are of greater concern to humans, from a health and environmental perspective, because they are respirable.

Along with size, the chemical constituent of PM influences its behaviour and the extent of its impact. For example, the toxicity of PM is determined by its configuration. Ingestion of toxic particles may produce a more harmful effect on the body than ingesting a more benignly composed particle (Bell et al., 2009).

1.3 Particulate matters influence on health, visibility and climate systems

Research has revealed more about the effects of emissions polluting the air we breathe. There is growing evidence of the serious health impacts and costs associated with air pollution (National Research Council, 2004, Kam, 2012). Additionally, air pollutions effect on visibility and climate systems is well established in the literature (Malm, 2000, Sloane et al., 1991).

Influence on Health

Perhaps the most serious consequence of high levels of PM in ambient air is its impact on human health. Air pollution, along with tobacco smoking and high blood pressure, are the three leading risk factors for global disease burden (Lim et al., 2013), with the World Health Organisation (2016) highlighting that approximately 3 million deaths per year are linked to exposure to outdoor air pollution. While air pollution in Australia is low when contrasted to other economically developed nations (Hansen et al., 2009), the population of Australia is most dense in major cities, where exposure to air pollution is omnipresent.

Exposure to ambient air pollution, in particular $PM_{2.5}$, is linked with serious negative effects on human health (National Research Council, 1998). The size of a particle is linked to its potential to be absorbed into the human body, with finer particles prompting a more severe impact on human health (Ferin et al., 1991, MacNee and Donaldson, 2003). The rate of particle deposition on the lungs for $PM_{2.5}$ is 50%, whereas coarse particles are usually removed in the nasal passage (Wark et al., 1998). Additionally, $PM_{2.5}$ has a larger surface area to mass ratio, with the consequence of them being more "biologically active" than coarser particulates (Kam, 2012, p.3, Oberdörster et al., 2005, Brown et al., 2001).

The resulting impact of PM on human health is an outcome of the period of exposure. Short-term exposure to increased air pollution can intensify existing respiratory and cardiovascular issues, along with increasing the chance of developing acute symptoms, hospitalisation and shortening lives (Australian Government Department of Environment and Energy, 2014, National Research Council, 2004, Haikerwal et al., 2015, Brook et al., 2010, Barnett et al., 2006, Barnett et al., 2005). Recurring long-term exposure can increase the likelihood of developing chronic respiratory and cardiovascular disease and mortality, can affect birth weight, and can cause irreversible effects to lung development in children (World Health Organization, 2013). Hansen et al. (2012) and Morgan et al. (2013) confirm this increase in morbidity and mortality associated with elevated $PM_{2.5}$ levels in an Australian context.

Influence on visibility

Another serious consequence of PM is its impact on visibility. Visibility is defined as the greatest distance which an object can be seen in a given direction with unaided eyesight (Wark et al., 1998). The degradation of visibility is attributed to fine particles in ambient air influencing the scattering and absorption of light that is transmitted through the atmosphere. The scattering of light is dependent on size, chemical composition and the hygroscopic nature of the particle, with fine PM being predominantly accountable for the reduction of visibility (Sloane et al., 1991, Malm, 2000).

Influence on climate systems

While high levels of PM in the air directly affect individuals in terms of how well they can see and the quality of the air they breathe, there are also consequences that cumulate beyond these immediate impacts. Some effects of PM are less direct and occur when increased aerosol concentrations from anthropogenic activities (mainly SO_2) produces increased concentrations of cloud condensation nuclei, leading to clouds possessing larger number concentrations of droplets with smaller radii, consequently leading to higher cloud albedos. Direct effects from aerosols can be observed as the sunlight that is reflected upward from a layer of haze. Aerosol particles cause a scattering of incoming solar radiation. This light scattering causes more solar radiation to be reflected from the earth back to space, ergo, a decrease in the amount of solar radiation reaches the earth's surface. This causes an overall cooling of the earth.

1.4 National Ambient Air Quality Standards

Due to the known adverse effects of air pollution on human health, visibility and climate systems, the Australian and State and Territory Governments agreed (through the National Environment Protection Council) to the National Environment and Protection Measure for Ambient Air Quality (AAQ NEPM), on the 26th of June, 1998. The goal of setting the AAQ NEPM is to protect health by defining the levels of PM, and other gaseous pollutants, in the air that should not be exceeded.

Six criteria pollutants were identified, due to their recognized negative effect on people, nature or materials, and national ambient air quality standards for each pollutant now exist. The pollutants include carbon monoxide, nitrogen dioxide, ozone, sulfur dioxide, lead and PM_{10}. As a result of ongoing research, it was recognized that smaller particles had great adverse health effects for humans. Hence, the AAQ NEPM was amended in 2003 to include reporting standards for $PM_{2.5}$. Lead monitoring ceased in 2004. The pollutant and their standards are shown in Table 1- 1. The NSW Government established a state wide air quality monitoring network to ensure compliance with these national goals.

Table 1- 1. National environmental protection standards for designated criteria pollutants set by the Australian Government. Source: National Environmental Protection Council (2015).

Pollutant	Averaging period	Maximum concentration standard
Carbon monoxide	8 hours	9.0 ppm
Nitrogen dioxide	1 hour	0.12 ppm
	1 year	0.03 ppm
Photochemical oxidants (as Ozone)	1 hour	0.10 ppm
	4 hours	0.08 ppm
Sulfur dioxide	1 hour	0.20 ppm
	1 day	0.08 ppm
	1 year	0.02 ppm
Lead	1 year	0.5 μg/m^3
Particles as PM_{10}	1 day	50 μg/m^3
Particles as $PM_{2.5}$	1 day	25 μg/m^3
	1 year	8 μg/m^3

1.5 Ambient monitoring

In order to comply with the AAQ NEPM standards, specific methods must be followed when measuring the concentration of pollutants. These are outlined in Schedule 3 of the AAQ NEPM, and are displayed in Table 1- 2. Such standards ensure a streamline approach to monitoring across the monitoring network. In Australia, the AAQ NEPM reference method for monitoring $PM_{2.5}$ is the manual gravimetric method. Continuous and automated methods can be employed as a substitute to the reference method.

Table 1- 2. Australian Standards Methods for $PM_{2.5}$ Pollutant Monitoring. Source: Federal Register of Legislative Instruments (2016).

Methods Title	Method Number
Determination of Suspended Particulate Matter-$PM_{2.5}$ low volume sampler-Gravimetric Method	AS/NZS 3580.9.10:2008
Determination of Suspended Particulate Matter-$PM_{2.5}$ beta attenuation monitors	AS/NZS 3580.9.12:2013
Determination of Suspended Particulate Matter-$PM_{2.5}$ continuous direct mass method using a tapered element oscillating microbalance monitor	AS/NZS 3580.9.13:2013
Determination of Suspended Particulate Matter-$PM_{2.5}$ high volume sampler with size selective inlet – Gravimetric Method	AS/NZS 3580.9.14:2013

1.6 Sampling methods

A range of methods are available for measuring PM concentrations in ambient air. Broadly, these can be classified as mass-only sampling or chemical speciation sampling. Both mass-only and chemical speciation methods are important, as the mass and the chemical composition of PM contributes to its impact on public health and the environment.

Mass-only sampling

Mass-only sampling typically requires collecting particles on filter paper and weighing the sample. It is expressed as the total mass of a particulate matter per unit volume based upon particle samples less than or equal to the specified aerodynamic diameter (Greene, 2005). The mass concentration of the PM is established, irrespective of its chemical composition.

A. Batch and continuous sampling

Mass-only sampling can be further categorized as batch or continuous sampling. Batch sampling involves sampling ambient air over a given time period, and then analysing this sample. This time period can be extensive, resulting in readings that are not in "real-time", often taking weeks to months before $PM_{2.5}$ data is available. This lag proves difficult for regulatory bodies to respond to significant air quality events. Conversely, continuous sampling methods record samples at much shorter intervals compared to batch samples, allowing for "real-time" data to be accessible for analyses. It is more advantageous to implement continuous sampling methods in a monitoring network for many reasons. Primarily, "real-time" data facilitates the decision-making process for regulatory agencies on their appointment of resources, while requiring little labor to operate, and providing more data for a low-cost. Continuous monitors can also assist with model development and validation, and source appointment, allowing regulatory agencies to monitor events that could be correlated to health effects (Chung et al., 2001).

B. Chemical speciation sampling

Another form of sampling is chemical speciation, which involves analyses to confirm the chemical composition of the PM. A range of techniques are available for this type of sampling, where the instrument utilized is dependent upon the constituent being evaluated. First, the sample is collected on the filter, then analysis techniques may include X-Ray Flourescence, Atomic Absorption Spectroscopy or Inductively Coupled Plasma Emission Spectroscopy. These types of instruments are very expensive and are less frequently used than mass-only devices. As highlighted by Chow (1995, p.326), the "chemical components found in an ambient air sample have a strong correspondence to the chemical composition of the source emissions in the monitored airshed". Chemical speciation allows for point sources of emission to be identified. Usually, chemical speciation methods are not in real-time, and therefore, do not offer information regarding the PM constituents promptly following sampling.

1.7 Mass-only sampling instruments

Two mass-only sampling instruments, both continuous samplers, were utilized in this study. They include the Beta Attenuation Monitor (BAM) 5014i and the Tapered Element Oscillating Microbalance (TEOM) 1400AB. The specifications of each instrument are provided in Appendix 1.

A. Beta Attenuation Monitor

Beta absorption was first utilized in the 1960's and 1970's as a technique to measure airborne PM (Husar, 1974, LilIenfeld, 1970). Since then, the instrument has advanced considerably. The Met-One BAM (5014i) is illustrated in Figure 1- 1. The instrument hosts a size selective inlet of 2.5 microns, along with filter tape, a beta radiation source, and a beta radiation detector. The BAM measures $PM_{2.5}$ mass by measuring the absorption of beta radiation by PM deposits on the filter tape. To account for blank attenuation, the attenuation is first measured on an unexposed section of tape. This section of tape is then exposed to ambient air for a given amount of time, while $PM_{2.5}$ is being deposited on the tape. The beta attenuation measurement is then performed again, and corrected using the blank attenuation measurement. Using this difference and the constant flow rate, the mass concentration is calculated. Continual monitoring is attained via an automatic mechanism that shifts the filter tape for each sampling event.

It has been demonstrated that relative humidity of ambient air significantly influences BAM readings, especially at high ambient relative humidities (Huang and Tai, 2008). Hence, Advanced Smart Heater technology is used in the instrument to precisely control the samples relative humidity. This aims to reduce particle bound water and to reduce positive artefact measurements thay may result due to condensation on the filter tape, or conditions of high humidity (Thermo Scientific, 2014). However, this too may bias particulate measurements when there is a large portion of volatile particulate matter present, as the heating drives off the volatiles (Chung et al., 2001).

B. Tapered Element Oscillating Microbalance

The Thermo Electron TEOM Series 1400AB Ambient Particulate Monitor was utilised in this study, as developed by Rupprecht & Patashnick Co., Inc. The device is illustrated in Figure 1- 2. The ambient air passes into the unit through an EPA standard PM_{10} size selecting sampling inlet. This inlet regulates the flow rate of 1 m^3/hr (16.7 L/min). When the sample stream leaves the inlet, the ambient air passes through a Very Sharp Cut Cyclone only allowing $PM_{2.5}$ to proceed through. Next, the 16.67 L/min flow is isokinetically split into a 3.0 L/min sample stream, where it is directed to the sensor unit. The TEOMs sensor

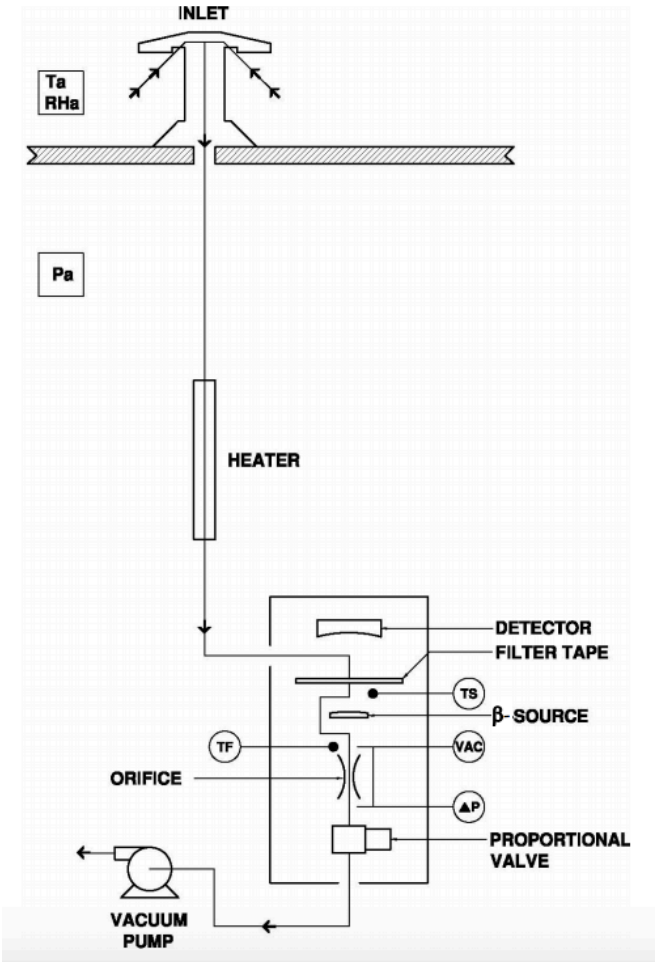


Figure 1- 1. Model of Beta attenuation monitor 5014i – flow schematic. Source: Thermo Scientific (2014).

unit contains a filter cartridge covering a hollow tapered glass element that oscillates when a parcel of air is drawn through the filter. The particles deposited on the filter alter the oscillation of the element, inversely proportionate to the particle mass. The mass and mass concentration can then be derived. The tapered element is reactive to small mass changes, and continuous monitoring in "real time" can be achieved.

Due to ambient air particles being hygroscopic, the TEOM heats the incoming air to 50 degrees Celsius under standard operating conditions, to prevent measurement issues associated with moisture or thermal expansion of the tapered element. Determining an accurate inlet tube temperature is crucial, as the measurement of $PM_{2.5}$ by the TEOM can be directly influenced by measuring particle bound water, or volatile compounds that are adsorbed on the PM (Greene, 2005, Charron et al., 2004, Rizzo et al., 2003, Eatough et al., 2003, Price et al., 2003).

As a result of heating to avoid collection of particle bound water, ammonium nitrate and semi-volatiles associated with fine particles are not retained on the collection filter,

meaning the TEOM only measures non-volatile PM (Long et al., 2002, Grover et al., 2005). What results is a reading that is not a true representation of the total ambient air concertation of PM. The manufacturer of the TEOM highlighted an issue with the device relating to volatization of the ambient air sample in 1993 (Rupprecht & Pataschnick, 1993). In 1997 Allen et al. (1997) outlined a varying relationship between the TEOM and the time-integrated gravimetric (manual) PM method. The degree of disparity is dependent on the monitoring location, time of year, and PM concentrations. Subsequently, many articles were published that evaluated the TEOM, concluding that the device provides unsatisfactory measurements of PM mass concentration in relation to traditional filter based methods, especially at low temperatures (Ayers et al., 1999, Rizzo et al., 2003, Charron et al., 2004, Li et al., 2012). The under-estimation of TEOM measurements is now well recognised.

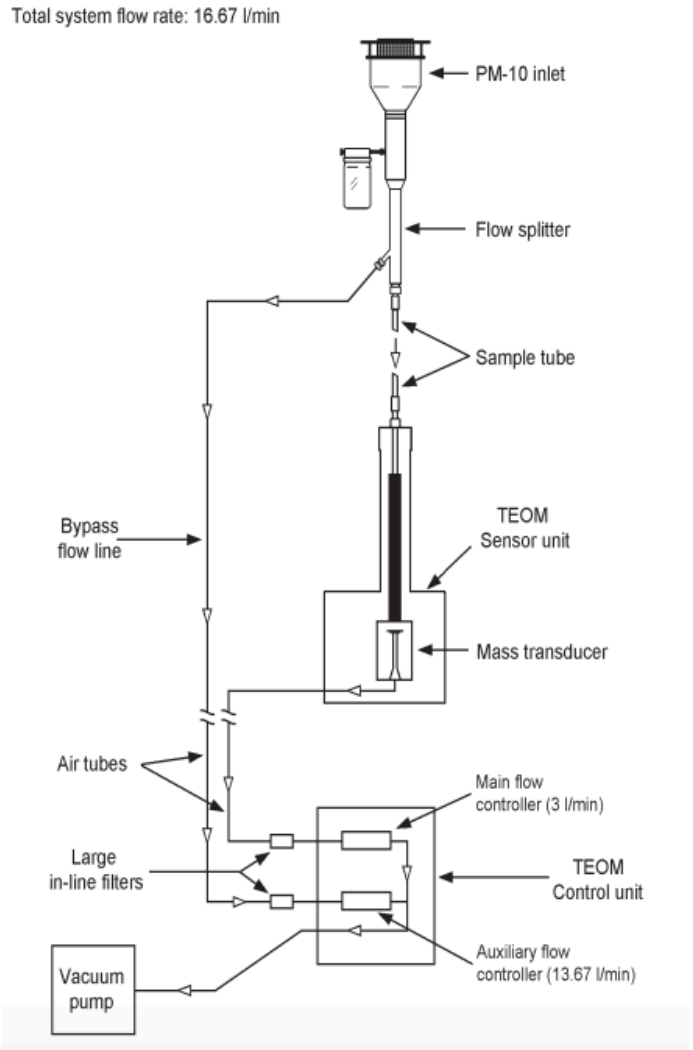


Figure 1- 2. Schematic diagram of flow for the TEOM1400AB. Source: Rupprecht & Pataschnick (2008).

1.8 Methods to resolve

There are a range of methods proposed in the literature that are used to correct TEOM measurements for their loss of semi-volatile material. A lot of these methods rely on having chemical speciation data, which is added to the raw PM values (Price et al., 2003, Charron et al., 2004, Chung et al., 2001, Hauck et al., 2004, Li et al., 2012, Godri et al., 2009). When the semi volatile material is accounted for, the agreement between instrument improves. However, we do not have chemical speciation data available for use.

Other methods explored apply simple correction factors to account for the underestimation of the TEOM instrument (Tsigaris, 2014). However, we know the TEOM $PM_{2.5}$ readings are influenced by other variables, including meteorological conditions and other air quality data. Therefore, a single correction factor is not suitable in our case.

Correction factors that incorporate air quality and meteorological covariates have been explored (Green, Fuller et al. 2001, Rizzo, Scheff et al. 2003, Gehrig, Hueglin et al. 2005, Winkel, Rubio et al. 2015), proving to be fairly successful.

Lastly, there are a range of different modelling techniques applied to correct $PM_{2.5}$ data, ranging from structural equation modelling (Bilonick et al., 2015), to non-linear regression (Kashuba and Scheff, 2008) and orthogonal regression (Hsu et al., 2016), all demonstrating promising results.

Perhaps the solution lies in a correction factor accounting for multiple variables or a prediction model. One cannot simply apply a correction technique/model developed on one data set to another data set, as our practical situation would possess a different underlying structure and would have departed from the ideal described by the assumptions made in the original model. It is for this reason that we did not have a set method in mind to apply to our data. Instead, an exploratory data analyses is performed, providing us with more of an insight into the data, guiding our decision to the most appropriate method to correct the TEOM data. The exploratory data analyses is performed in Chapter 3. A more thorough review of the literature is provided in Appendix 2.

From here onwards, $PM_{2.5}$ TEOM will simply be referred to as TEOM. And PM_{10} TEOM will be referred to simply as PM_{10}.

Chapter 2: $PM_{2.5}$ in Sydney

2.1 Influence on air quality

The data used in this study is sourced from the air quality monitoring station at Chullora in Sydney, Australia, operated by the Office of Environment and Heritage (OEH). Parameters measured here include ozone, carbon monoxide, sulfur dioxide, nitrogen oxide, nitrogen dioxide, nitrogen oxides, fine particles (by nephelometry), fine particles ($PM_{2.5}$ using a TEOM and BAM, and PM_{10} using a TEOM), wind speed, wind direction, ambient temperature and relative humidity. Apart from differing emission sources and their strengths, PM concentrations can be affected by local topography, climate, meteorological conditions and secondary chemical reactions (Crawford et al., 2016a, Davis and Gay, 1993, Beaver et al., 2010), especially in a confined air-shed like the Sydney basin (Crawford et al., 2016b).

Topography

$PM_{2.5}$ samplers were collocated at Chullora; located at 33 ° 53’ 38”S, 151 ° 02’ 43” E, 10 metres above sea level. Chullora lies in the greater urban area of the Sydney. The Sydney basin is approximately 200 km north-south and 100 km east-west, surrounded by the Great Dividing Range to the west, which runs parallel to the east coast and is approximately 1 km above sea level. Mounts approximately 200 metres above sea level border the north and south of the basin. Chullora is located approximately in the centre of the Sydney basin in a mixed urban and residential area. Cohen et al. (2011) and Cohen et al. (2012) note that the Sydney basin acts as a trap for fine particle pollution that is generated locally, and particle pollution that is transported into the basin from external sources, like soil and desert dust (Leslie and Speer, 2006).

Climate

The concentrations of PM measured at one particular site are known to be influenced by the local meteorological conditions, chemical transformation, and synoptic weather systems (Crawford et al., 2016b, Jiang et al., 2005). There is a relationship between synoptic weather systems and PM, with high pressure systems resulting in high PM concentrations (Huang et al., 2009, Crawford et al., 2016b). Typical conditions accompanied by a high pressure system include low wind speeds and a low rate of pollutant dispersion. However it has also been demonstrated that in some cases, low pressure systems can be a catalyst for high PM concentrations, due to the strong winds stirring up PM in the form of soil dust (Dayan and Levy, 2005). During the winter months, inversions can trap and concentrate pollution in the basin, whereas the summer months are normally accompanied by a sea breeze that pushes pollution inland from the coastal region. Figure 2- 1 illustrates a wind rose for the

Chullora area during the study period, from 02/09/2010 to 29/11/2012. The wind velocity provides a measure of the mean transport direction and pollutant ventilation.

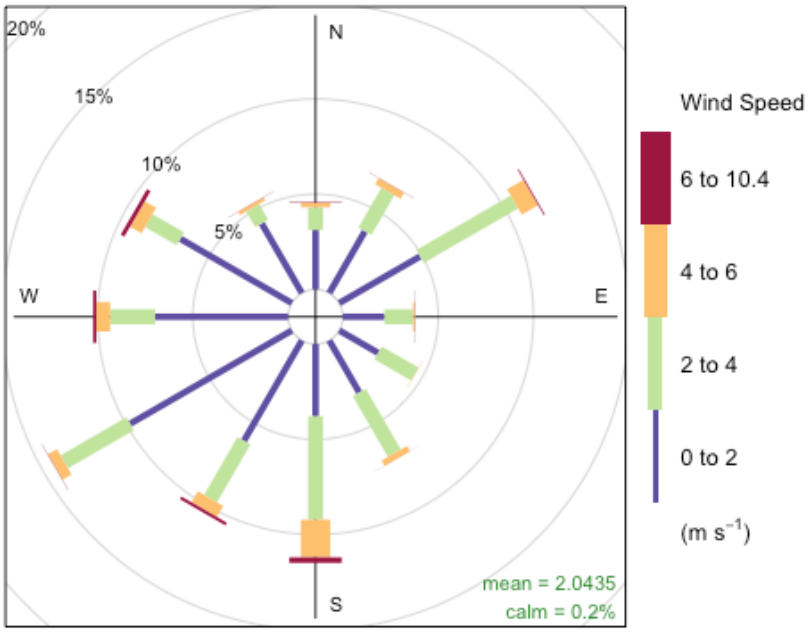


Figure 2- 1. Wind Rose: Frequency of counts by wind direction (%) for Chullora during the study period.

Australia is one of the driest continents, home to a dozen desert regions covering approximately 18% of total land mass, situated mainly in the central and north western areas of the country. The rainfall in these regions can be as less as 100mm/year, with temperatures above 40 degrees Celsius for long periods of time, causing high evaporation rates resulting in severe soil moisture deficits and reduced vegetation cover. The combination of these conditions produces approximately 5 to 10 significant dust storm events yearly, which can significantly impact the Sydney area, by reducing visibility (Ekström et al., 2004) and increasing aerosol loading (Mitchell et al., 2010). Average temperature and precipitation patterns for Chullora during the study period are shown in Table 2- 1. The rainfall data was recorded at the closest rain gauge at Strathfield Golf Club, located approximately 3km away from the Chullora site. The ambient temperature was recorded on site.

Table 2- 1. Average temperature and precipitation for period of collocation of BAM and TEOM instruments at Chullora site, between 02/09/2010 & 29/11/2012.

Month/Year	Average ambient temperature (° C)	Average Precipitation (mm)
January	23.15	65.5
February	22.85	60.0
March	20.86	191.0
April	18.01	174.5
May	13.74	58.5
June	12.32	118.0
July	11.43	97.50
August	13.60	26.50
September	15.32	62.67
October	17.07	47.0
November	19.89	141.0
December	19.77	72.0

There is often significant seasonal variation in $PM_{2.5}$ (Allen et al., 1997) due to a range of factors including; meteorology, power production from combustion sources, solar radiation available, and other factors relating to the formation of secondary $PM_{2.5}$ (Greene, 2005). Seasonal $PM_{2.5}$ variation for Chullora is shown in Figure 2- 2. TEOM $PM_{2.5}$ concentrations tend to be higher during warmer months (spring and summer), with the exception of the month December, and lower during cooler months (autumn and winter), with the exception of the month March. The BAM tends to follow the same pattern as the TEOM, except for in the summer months. The $PM_{2.5}$ concentrations do not drop as low in December for the BAM, and are fairly constant in January and February.

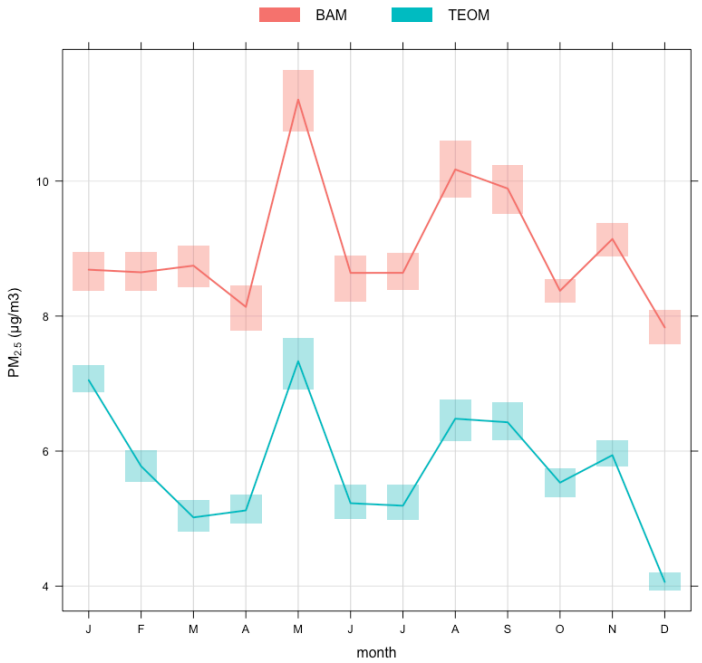


Figure 2- 2. $PM_{2.5}$ concentrations illustrating the seasonality of the TEOM and BAM data for the collocated period.

2.2 Sources and chemical contribution of $PM_{2.5}$

PM can be present in ambient air as a result of primary or secondary processes. Its chemical composition is a result of the source of the particles and any chemical alterations that occur within the particle. Hence, the source of the PM can be inferred by determining the chemical composition of the PM.

Secondary formation of $PM_{2.5}$ is a physiochemical process, making it difficult to determine sources of PM exactly, especially when many sources contribute to the composition of the PM present. When investigating the origin of secondary $PM_{2.5}$, a high level of uncertainty is introduced, as the precursor gas emitters, wind patterns, residence times and removal times must all be known and accounted for (Greene, 2005). It is further complicated by the fact that the sources of $PM_{2.5}$ may not exist in the area surrounding the receptor site. Primary sources of $PM_{2.5}$ are more simple to trace, and can include pollution from agriculture, roads, domestic wood combustion, forest fires, fugitive dust, and industry.

The Sydney Fine Particle Study (Cohen et al., 2016) applied positive matrix factorization source appointment methods on daily $PM_{2.5}$ data from 4 sites in the greater Sydney region (from 2000 to 2014), to identify elemental composition sources and to quantify their contribution to the total $PM_{2.5}$ at each site. The sampling sites investigated in the study were Liverpool, Lucas Heights, Mascot and Richmond, which are located 20km, 21km, 22km and 53km respectively from Chullora. Across all sites, the average $PM_{2.5}$ concentration was 6.82 $\mu g/m^3$. This was divided into seven source fingerprints, as summarized in Table 2- 2. Although these averages are based on data from the year 2000 to 2014, from sites that are at least 20km from Chullora, they still provide a good indication of the possible $PM_{2.5}$ mass loadings on a broader scale.

The results from the Particle Study reveal that mixed secondary sulfate and mixed aged industrial sulfate fingerprints made up 50-70% of $PM_{2.5}$ in summer, while smoke from biomass burning contributed 60-80% to total $PM_{2.5}$ concentrations in winter, as a result of domestic wood combustion.

One of the largest $PM_{2.5}$ contributors in Sydney is ammonium sulfate (ANSTO, 2010, Cohen et al., 2012, Cohen et al., 2016). In Sydney the sulfate component is fairly consistent on a spatial scale, however it is strongly influenced by season (ANSTO, 2010). Ammonium sulfate concentrations are twice as high in summer than in winter, possibly due to intensification of photochemical activity and higher energy demand (Chan et al., 2008), along with sunlight, UV, temperature and humidity all facilitating its formation. Five coal-fired power stations are currently operating in New South Wales (Bayswater: 2,640 MW, Liddell: 2,000 MW, Mt Piper: 1,400 MW, Eraring: 2,880 MW and Vales Point: 1,320 MW), although

eight were operating during the study period (Munmorah: 600 MW (closed in 2012), Wallerawang (1,000 MW (closed in 2014), Redbank: 151 MW (closed in 2014)). While located many kilometres away from the Sydney metropolitan area, they still contribute significantly to the fine particle mass is Sydney. In 2011, up to half of the total sulfate air pollution, and 18% of the total $PM_{2.5}$ in the greater Sydney region was caused by emissions from these eight coal-fired power stations (Cohen et al., 2011).

Table 2- 2. Average $PM_{2.5}$ source fingerprints across Liverpool, Lucas Heights, Mascot and Richmond between 2000 and 2014. Source: (Cohen et al., 2016).

Source Fingerprint	Average $PM_{2.5}$ mass	%	Description
Soil	0.25	(4+-5)%	Represents fine wind-blown dust
Seam	0.51	(10+-11)%	Represents sea spray transported from the coast.
Mixed-2ndryS	1.63	(24+-16)%	Represents secondary sulfates, indicative of coal power stations, oil refineries, motor vehicles and industry.
Mixed-Ind-Saged	0.95	(15+-13)%	Represents industrial sources featuring components of aged secondary sulfates and sea spray.
Mixed-smoke-auto	2.08	(24+-20)%	Represents smoke from biomass burning, especially from domestic wood heaters in the winter with components from diesel vehicles.
Auto1	1.22	(20+-10)%	Represents the automobile.
Auto2	0.23	(3+-2)%	Represents a second minor automobile source, associated with the use of leaded petrol which ceased in 2001.

During summer 2011 and autumn 2012, Cope et al. (2014) analysed the percent contribution of chemical source groups to the $PM_{2.5}$ mass concentration in Westmead, Sydney, located approximately 15 kilometres SSE of Chullora. The Summer 2011 program classified sea salt (34%) and organic matter (primary and secondary; 34%) as the major contributors to the composition of $PM_{2.5}$, with secondary inorganic aerosol (15%), soil (11%) and elemental carbon (6%) also contributing to the make-up of $PM_{2.5}$. Further isotopic analysis of the organic matter reveals that up to 70% of the analysed carbon is modern (Cope et al., 2014). Hence, secondary organic aerosols are formed through biogenic sources. The autumn 2012 program displays a reduced sea salt contribution (5%) and an increase in organic matter contribution (57%). The elemental carbon also increased (16%), with soil

decreasing (7%) and secondary organic aerosols remaining the same (15%). The percent contributions for the summer 2011 and autumn 2012 period are also shown in Figure 2- 3.

Given Sydney has a population of over 5 million people, and approximately 3 million motor vehicles are in operation in the Sydney basin, one would anticipate that the majority of $PM_{2.5}$ would be produced within the basin. However, this is not necessarily the case. Observational and modeling studies reveal that aerosol concentrations are affected by long-range transport, in combination with anthropogenic and natural emissions (Jacob et al., 2003, Jaffe et al., 2003, Liu and Mauzerall, 2005, Kan et al., 2007). We can conclude that there are a number of sources affecting the PM measurements at the Chullora site.

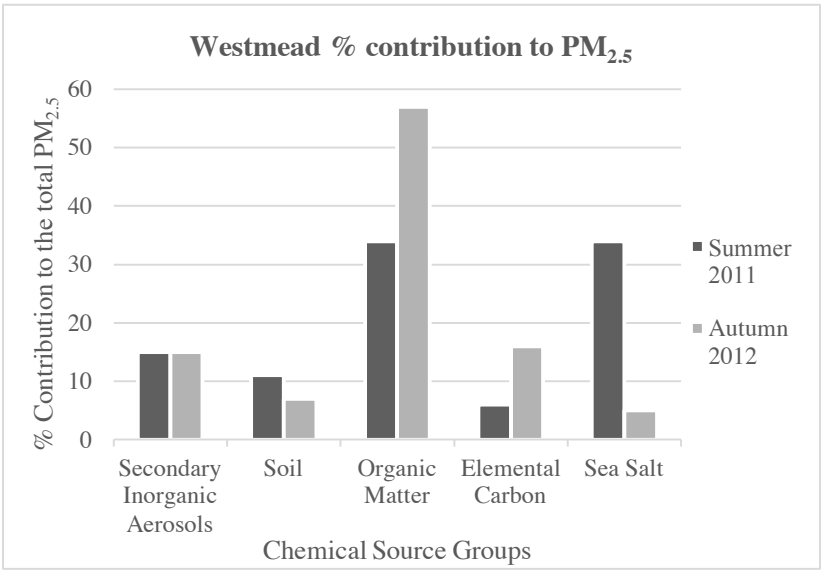


Figure 2- 3. Percentage contribution of chemical source groups to Summer 2011 and Autumn 2012 average $PM_{2.5}$ concentrations. Source: Cope et al. (2014).

2.3 Volatiles

As previously mentioned, atmospheric aerosols are composed of a range of species. Significant components can include trace metals, crustal materials, elemental carbon and sulfate, which are deemed to be stable species that can be precisely measured (Musick, 1999, Salvador and Chou, 2014). However, semi-volatile materials, which also make a substantial contribution to the mass loading of $PM_{2.5}$ (Lewtas et al., 2001, Tang et al., 1994) are unstable, existing in both the gas and particulate phase. Semi-volatile material may include hygroscopic material (particle bound water), semi-volatile organic compounds and ammonium nitrate in equilibrium with nitric acid and ammonia. It is widely acknowledged that the TEOM has the shortcoming of driving off semi-volatile material (Allen et al., 1997, Charron et al., 2004, Cyrys et al., 2001). As the deposited mass of the TEOM has to be

heated to a temperature above ambient levels, water and most of the semi-volatiles are evaporated. As a result, the TEOM substantially underestimates the $PM_{2.5}$. These semi-volatile components can be more accurately measured using other samplers, like the BAM.

Studies of aerosols in urban environments have demonstrated that a considerable portion of $PM_{2.5}$ consists of semi-volatile organic and nitrate materials (Long et al., 2002, Long et al., 2003, Eatough et al., 2001, Grover, 2006). Hence, there is a great deal of uncertainty in measurement recorded using a TEOM instrument. Additionally, semi-volatile fine particulate organic material tends to be secondary in nature (Eatough et al., 2003, Long et al., 2002), making it difficult to determine the mechanisms and kinetics of the formation of these particles (Grover, 2006).

2.4 Monitoring and management in Sydney

The OEH operates monitoring stations all over New South Wales. Their remit includes monitoring and analysing air quality and providing this vital information to the community and industries. The OEH then works with the Environmental Protection Authority and NSW Health to develop ways to reduce air pollution and to protect the health and well-being of the community.

Chapter 3: Exploratory data analyses

3.1 Overview

Exploratory data analysis offers conceptual and computational tools for identifying patterns in data to assist with hypothesis development and refinement. It postpones the usual assumption about what type of model the data will follow, allowing a more direct approach of letting the data itself reveal its underlying structure, and consequently reveal the most appropriate model to be applied to the data.

The program R (www.r-project.org) is an open source programming language and software environment that is widely used for statistical computing and graphics across many disciplines (R Development Core Team, 2011). It offers exceptional interactive analysis capabilities, and is suitable for efficient development of statistical and data analysis applications, like exploratory data analyses. *Openair* is an R package built for the purpose of analysing air pollution measurement data, and more broadly to be applied to the atmospheric sciences (Carslaw and Ropkins, 2012). *Openair* was used in our study due to its suitability for our analysis.

3.2 Available data

Data for analysis was sourced from the OEH's air quality monitoring site located at Chullora. The suite of air pollution indicators and meteorological parameters measured at this site include: $PM_{2.5}$ (using a TEOM and BAM), ambient temperature, relative humidity, scattering of light by fine particles (using a nephelometer), carbon monoxide, oxides of nitrogen (NO_x, NO_2, NO), sulfur dioxide, ozone, PM_{10} (using a TEOM), wind direction, variation in wind direction, and wind speed (Table 3- 1). The span of the period of collocation of the TEOM and BAM at Chullora was from 02/09/2010 at 5:00 p.m. to 29/11/2012 at 11:00 a.m.

The OEH employs quality assurance procedures for air quality monitoring in the Sydney network, meaning their data is precise (through daily calibration checks), accurate (through multi-point calibration), representative and comparable to other institutions using similar methods (Office of Environment & Heritage, 2015). Any instrument that is not operating correctly automatically has its data invalidated. We then assume that any value recorded is correct, and exists because it is valid. Therefore, as little data cleaning was performed as possible.

Table 3- 1. Descriptive statistics for air pollution and meteorological parameters, shown seasonally, based on hourly data.

Parameter	Season	Variance	Minimum Value	Maximum Value	Standard Deviation	Mean
$PM_{2.5}$ (BAM) ($\mu g/m^3$)	Autumn	55.3	-2.5	62.9	7.4	9.5
	Spring	45.9	-2.5	121.6	6.8	9.1
	Summer	31.0	-2.5	39.6	5.6	8.4
	Winter	41.4	-2.4	46.6	6.4	9.1
$PM_{2.5}$ (TEOM) ($\mu g/m^3$)	Autumn	32.8	-2.4	49.1	5.7	5.9
	Spring	35.2	-2.5	170.9	5.9	6.0
	Summer	16.9	-2.4	44.9	4.1	5.8
	Winter	28.4	-2.4	65.3	5.3	5.7
PM_{10} (TEOM) ($\mu g/m^3$)	Autumn	230.2	-4.6	366.5	15.2	18.8
	Spring	166.5	-8.4	361.6	12.9	19.3
	Summer	159.3	-5.4	236.2	12.6	18.5
	Winter	326.4	-4.3	386.8	18.1	18.9
Nephelometer (bsp)	Autumn	0.1	0.01	3.11	0.3	0.3
	Spring	0.1	0.01	12.25	0.3	0.3
	Summer	0.0	0.01	1.52	0.1	0.2
	Winter	0.1	0.01	4.02	0.3	0.3
Temperature (°C)	Autumn	22.5	4.1	33.3	4.7	17.4
	Spring	24.6	4	36.9	5.0	17.4
	Summer	17.0	12.4	41.3	4.1	22.4
	Winter	13.7	2.7	28.4	3.7	12.3
Relative Humidity (%)	Autumn	316.6	17.1	100	17.8	73.4
	Spring	382.5	12.5	99.7	19.6	66.7
	Summer	274.6	14.2	98.9	16.6	71.7
	Winter	373.7	21.9	99.9	19.3	70.8
Carbon Monoxide (ppm)	Autumn	0.1	0.1	2.7	0.3	0.4
	Spring	0.0	-0.1	2.1	0.2	0.3
	Summer	0.0	-0.1	1.1	0.1	0.3
	Winter	0.1	0	3.5	0.3	0.4
Nitrogen monoxide (ppb)	Autumn	1174.6	-2	323	34.3	19.2
	Spring	360.0	-2	197	19.0	9.4
	Summer	171.0	-1	199	13.1	6.1
	Winter	1420.3	-2	432	37.7	23.0
Nitrogen Oxides (ppb)	Autumn	1544.3	1	359	39.3	33.3
	Spring	609.3	-2	226	24.7	22.9
	Summer	259.2	0	221	16.1	15.0
	Winter	1862.8	1	470	43.2	39.4
Nitrogen Dioxide (ppb)	Autumn	57.4	1	48	7.6	14.0
	Spring	70.2	0	56	8.4	13.4
	Summer	27.3	1	38	5.2	8.8
	Winter	67.1	1	51	8.2	16.3
Sulfur Dioxide (ppb)	Autumn	1.7	-1	25	1.3	0.7
	Spring	1.5	-2	15	1.2	0.7
	Summer	1.7	-1	24	1.3	0.7
	Winter	1.8	-2	26	1.4	0.7
Ozone (ppb)	Autumn	89.1	0	57	9.4	10.6
	Spring	134.7	-1	77	11.6	16.9
	Summer	122.7	0	99	11.1	14.0
	Winter	80.5	-1	40	9.0	10.4
Wind Speed (m/s)	Autumn	1.4	0	7	1.2	1.9
	Spring	1.8	0	10.4	1.3	2.1
	Summer	1.7	0	8.8	1.3	2.2
	Winter	1.7	0	9.2	1.3	2.0
Wind Direction (°)	Autumn	7635.4	0	360	87.4	201.4
	Spring	9824.6	0	359.8	99.1	173.6
	Summer	8321.0	0.1	359.8	91.2	151.8
	Winter	6458.9	0	360	80.4	230.4

The TEOM and BAM concentrations were reported hourly in micrograms/cubic meter ($\mu g/m^3$) by the OEH. We averaged the hourly concentrations over the 24-hour (1:00 a.m. to midnight) period in line with the national air quality guidelines and protocols (Office of Environment & Heritage, 2012). That is, days with less than 75% data capture are excluded from the 24-hour averages. Table 3- 2 shows the number and percent of missing data for all variables for the hourly and daily averages, for a total of 19,651 observations for the hourly data and 819 observations for the daily data.

Table 3- 2. Number of and percentage of missing data for all variables available for the study recorded over the collocated period, for hourly and daily averages.

Parameter	Hourly averaged data (19,651 total possible observations)		Daily averaged data (819 total possible observations)	
	No. of missing values	% of data missing	No. of missing values	% of data missing
$PM_{2.5}$ (BAM) ($\mu g/m^3$)	992	5.05	36	4.40
$PM_{2.5}$ (TEOM) ($\mu g/m^3$)	260	1.32	6	0.73
PM_{10} (TEOM) ($\mu g/m^3$)	148	0.75	5	0.61
Nephelometer (bsp)	46	0.23	1	0.12
Temperature (°C)	32	0.16	1	0.12
Relative Humidity (%)	32	0.16	1	0.12
Carbon Monoxide (ppm)	1200	6.11	15	1.83
Nirogen monoxide (ppb)	2134	10.86	20	2.44
Nitrogen Oxides (ppb)	1309	6.66	20	2.44
Nitrogen Dioxide (ppb)	1305	6.64	20	2.44
Sulfur Dioxide (ppb)	1369	6.97	24	2.93
Ozone (ppb)	1139	5.80	12	1.47
Wind Speed (m/s)	89	0.45	1	0.12
Wind Direction (°)	89	0.45	1	0.12

3.3 Comparisons of measurements from the collocated TEOM and BAM

The BAM and TEOM differ in their mean, with BAM possessing higher readings than the TEOM for all seasons (Table 3- 1). There is also a greater variance of BAM than TEOM in all seasons (Table 3- 1).The minimum value for both instruments is negative (Table 3- 1).

A scatterplot of the hourly TEOM and BAM readings for the collocated period is shown in Figure 3- 1A). These two sampling methods do not agree terribly well, as indicated by the R^2 value of 0.38 and the cloud of points in the bottom left corner of the plot, with a lot of scatter either side of the regression line (Figure 3- 1A). Most of the data lies between 0 and 25$\mu g/m^3$ for the hourly data, as indicated by red rings showing the higher density of points

(Figure 3- 1 A). The standard error for the intercept and slope coefficient are 0.06 and 0.01 respectively. The *y*-intercept of 4.62 $\mu g/m^3$ may be indicative of a systematic offset between the two methods, but this number alone cannot determine if the BAM is biased high or the TEOM is biased low, or a combination of both. Alternatively, it may be due to some outliers having a large influence on this line.

Figure 3- 1 B depicts the average daily TEOM and BAM readings for the collocated period. When averaged daily, there is good agreement between the TEOM and BAM, with an R^2 of 0.80. There is some scatter along the line, but a lot less than the hourly scatterplot (see Figure 3- 1A). Additionally, the cluster of data from the bottom left corner has drastically reduced. the majority of the data is bound by 0-10 $\mu g/m^3$ and 0-15 $\mu g/m^3$ on the x and y axis respectively (Figure 3- 1 B).The standard error for the intercept and slope coefficient are 0.13 and 0.02 respectively. Again, the *y*-intercept of 2.62 may indicate a systematic offset between the two methods, but this alone cannot determine if the BAM is biased high or the TEOM is biased low, or a combination of both.

A quantile-quantile plot (Q-Q plot) was used to determine if the two data sets of TEOM and BAM data come from populations within a common distribution. A Q-Q plot shows the quantiles of the first data set matched with the quantiles of the second data set. Figure 3- 2 B) is a magnification of 3A, with the *x* and *y* limits set to 50 $\mu g/m^3$, to help show the deviance from the reference line. The TEOM and BAM do not come from populations within a common distribution, as the points do not follow the 45-degree reference line (Figure 3- 2 A). The BAM values are biased higher than the corresponding TEOM values. The difference in these readings remains fairly constant between 10 and 50 $\mu g/m^3$. Interestingly, beyond 60$\mu g/m^3$ on the *x* and *y* axis, the devices read similar results, as the points are falling on the 45-degree line, up until ~160 $\mu g/m^3$ on the *x*-axis, where the TEOM gives a significantly higher reading for one paired sample.

Boxplots are used to compare key features of the BAM and TEOM distributions (Figure 3- 3). The box centerline illustrates the median, with the upper box limit representing the 75th percentile, and the lower box limit showing the 25th percentile. The whiskers on these plots reach a point equal to the range multiplied by the interquartile range. Outliers are not shown on these plots (Figure 3- 3). There is a clear difference between the BAM and TEOM measurements, with the TEOM having lower median value (4.60 $\mu g/m^3$), while BAM records a higher median value (8.00 $\mu g/m^3$) (Figure 3- 3). The BAM boxplot also displays more variation in the recorded measurements, since it has a wider inter-quartile range (Figure 3- 3). Summary statistic for TEOM and BAM are shown in Table 3- 3.

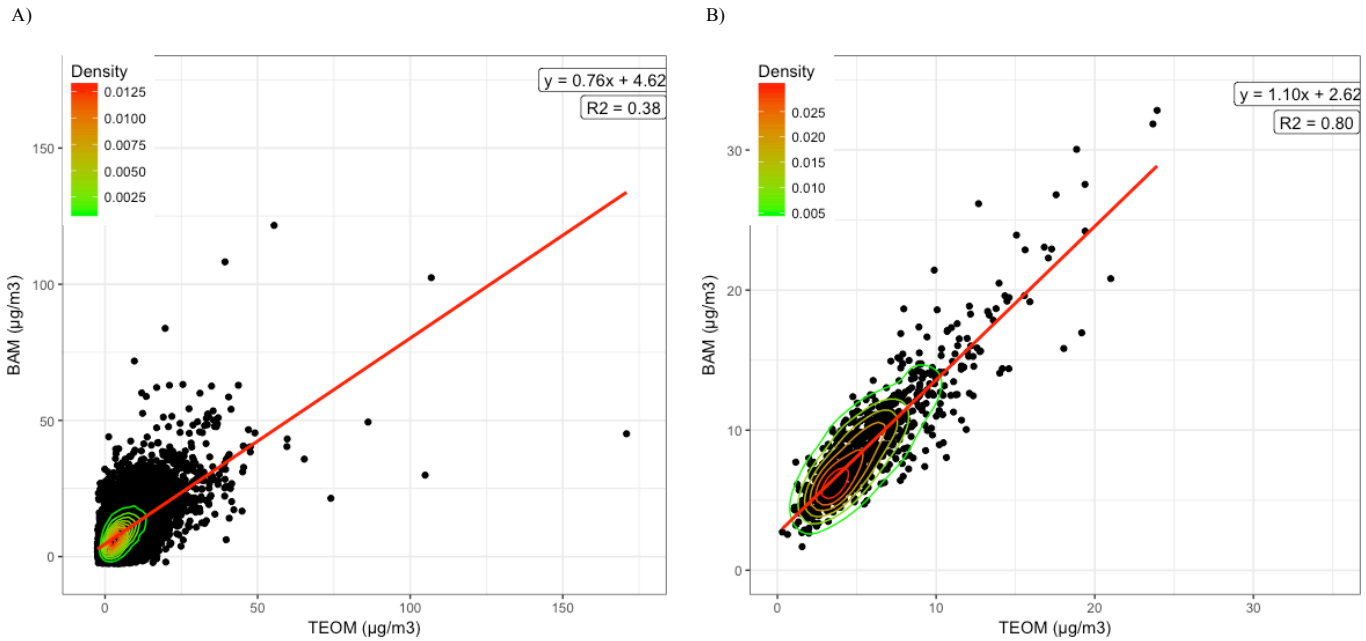


Figure 3- 1. Scatter plot showing density of points for the TEOM and BAM, A) showing hourly measurements and B) showing daily measurements for the collocated period. The fitted ordinary least squares regression line, R^2 and the coefficients are also shown.

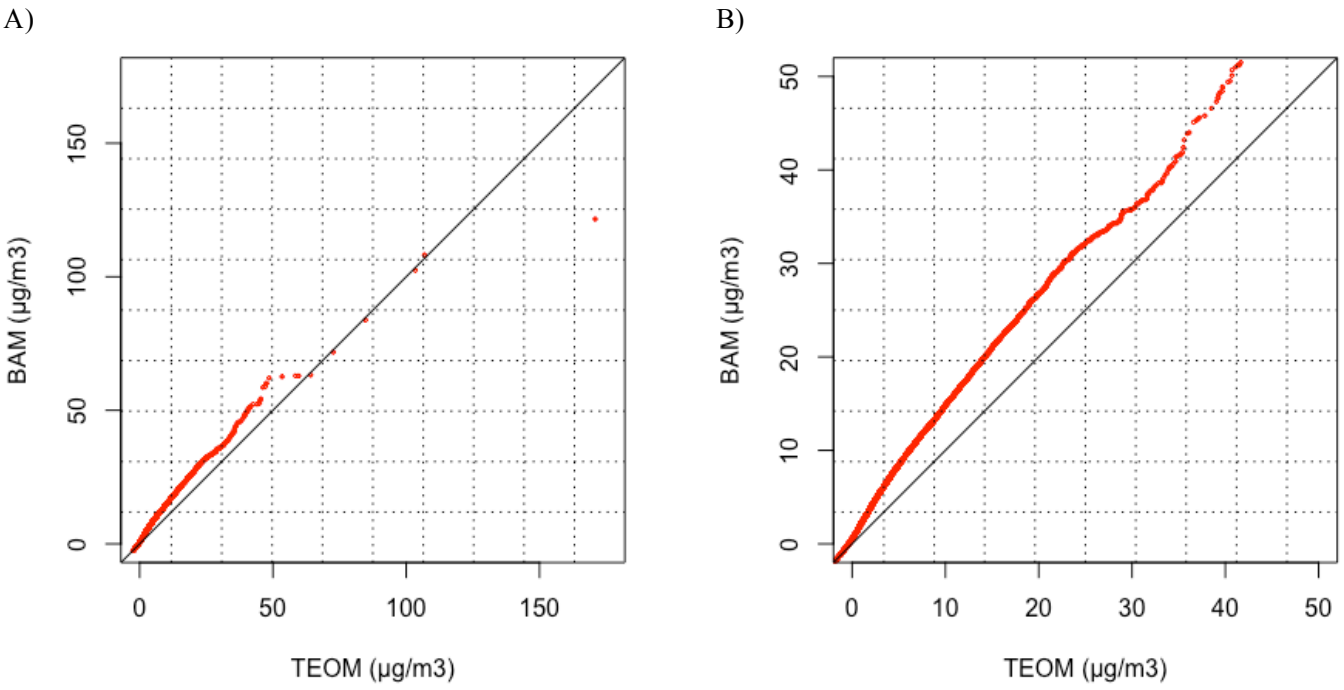


Figure 3- 2. Quantile-quantile plot of TEOM and BAM hourly measurements during the collocated period. A) shows the Q-Q plot for all data. B) shows the Q-Q plot with the *x* and *y* limits set to minimum 0 µg/m^3 and maximum 50 µg/m^3.

Table 3- 3. Instrument inter-comparison through basic statistics.

	TEOM (µg/m^3)	BAM (µg/m^3)
Mean	5.78	9.05
Median	4.60	8.00
Standard Deviation	5.33	6.62
Inter quartile range	5.20	7.10

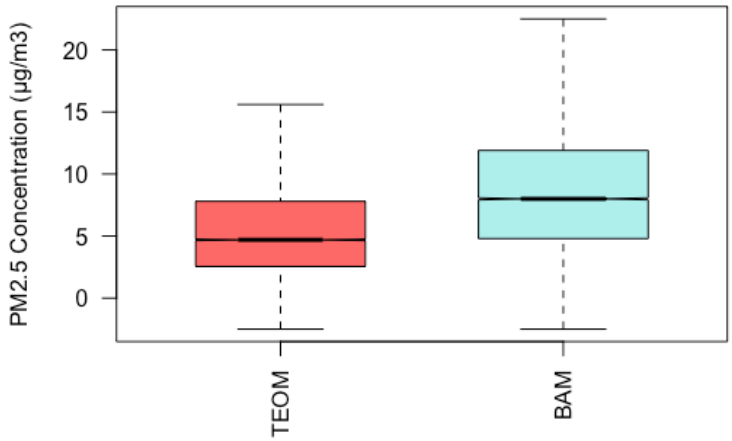


Figure 3- 3. Box and whisker plot showing BAM and TEOM measurements, based on hourly averages, for the collocated period (outliers excluded from display).

Density histograms (see Figure 3- 4) demonstrate that the data is skewed right, as depicted by the longer tail of the distribution on the right hand side than on the left hand side of both histograms. Additionally, the fact that the mean is greater than the median for both histograms (TEOM: median = 4.60 µg/m^3, mean = 5.78 µg/m^3 ; BAM: median = 8.0 µg/m^3, mean = 9.05 µg/m^3) indicates that the data is skewed, not showing a normal distribution. It is typical that air quality data is not normally distributed (Bouis, 1999, Nelson, 1980), as in principle, the lower limit of $PM_{2.5}$ never falls below zero, and the maximum value can have very high values, far from the mean. However, Kahn (1973) explains that air pollutant data has a lognormal distribution, suggesting that a log transformation may be appropriate to normalize the data in this case.

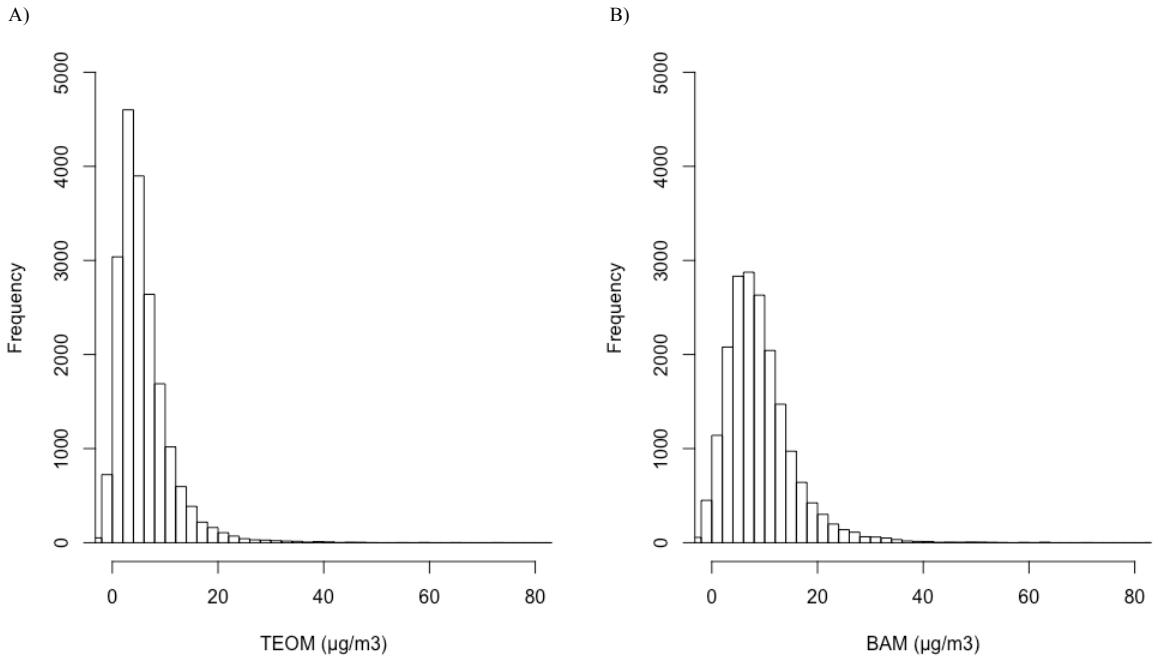


Figure 3- 4. Frequency histogram of data for the A) TEOM and B) BAM instruments, based on hourly averages, for the collocated period. The *x*-limit for both plots was set to 80 µg/m^3, There are 4 values each for the TEOM and BAM that were cut off in this plot as they are greater than 80 µg/m^3.

Having explored the distribution of the $PM_{2.5}$ data in detail, we now examine the hourly, daily and monthly variations in $PM_{2.5}$, by using time variation plots. We can observe which months have the highest and lowest mean concentrations, determine differences in instrument measurements, and can observe hourly, daily and monthly trends. The time variation plots show the mean, with the shaded colours illustrating the 95% confidence interval in the mean.

Exploring data at an hourly interval enables analysis at a fine scale. The TEOM increases from its low just before 6:00 a.m. and continues rising until it reaches approximately 9.00 µg/m^3 at around 8:00 a.m., during peak hour traffic (Figure 3- 5 B). The TEOM approaches the BAM at this time. Afterwards, the TEOM decreases to its lowest reading of the day at around midday, with a value ranging between 3.00 µg/m^3 and 4.00 µg/m^3 (Figure 3- 5 A). This then rises to approximately 7.00 µg/m^3, presumably as a result of afternoon traffic. On the other hand, the BAM reads higher than the TEOM between midnight and around 6:00 a.m. (Figure 3- 5 B). The BAM and TEOM readings follow each other quite well during the day, maintaining a fairly constant difference, between 6:00 a.m. and midnight, with the afternoon peak for the BAM readings around 10 µg/m^3 (Figure 3- 5

B). The largest discrepancy in the readings occurs at night time, around 2:00 a.m. till 4:00 a.m (Figure 3- 5 B).

Differences in the instrument measurements are also examined for each day of the week. Saturday and Sunday do produce lower measurements of $PM_{2.5}$ readings than weekdays (Figure 3- 5 D). This may be due to less cars being on the road due to less cars commuting for work. The difference between the readings remains fairly constant from Monday to Sunday. The peak on Tuesday (Figure 3- 5 D) could be due to the two highest TEOM values and four highest BAM values being recorded on the 04/09/2012, which was a Tuesday.

The pattern of monthly averages of the TEOM and BAM follow reasonably well, from March through to December (Figure 3- 5 C). The differences between the reading for these months is approximately 3.50 $\mu g/m^3$. However, in January and February the TEOM and BAM readings are much closer, with a difference of approximately 1.80 $\mu g/m^3$ and 2.50 $\mu g/m^3$ respectively.

Seasonal difference in the measurements are also observed. The difference in the measurements in this morning period between the instruments is exacerbated especially in winter, slightly lesser in autumn and spring (Figure 3- 6). TEOM readings are still lower than BAM during 2:00am and 4:00am in summer too, but not to the same extent (Figure 3- 6). During summer and autumn, the BAM and TEOM read quite closely at 9:00 p.m (Figure 3- 6).

The monthly averages for the collocated period, broken up by year, is shown in Figure 3- 7. In the warmer months, December to March, the $PM_{2.5}$ BAM levels differ markedly from 2010/2011 compared to 2011/2012. This highlights the complexity of the data, and the high level of difficulty there will be in building a model that can capture these changes in $PM_{2.5}$.

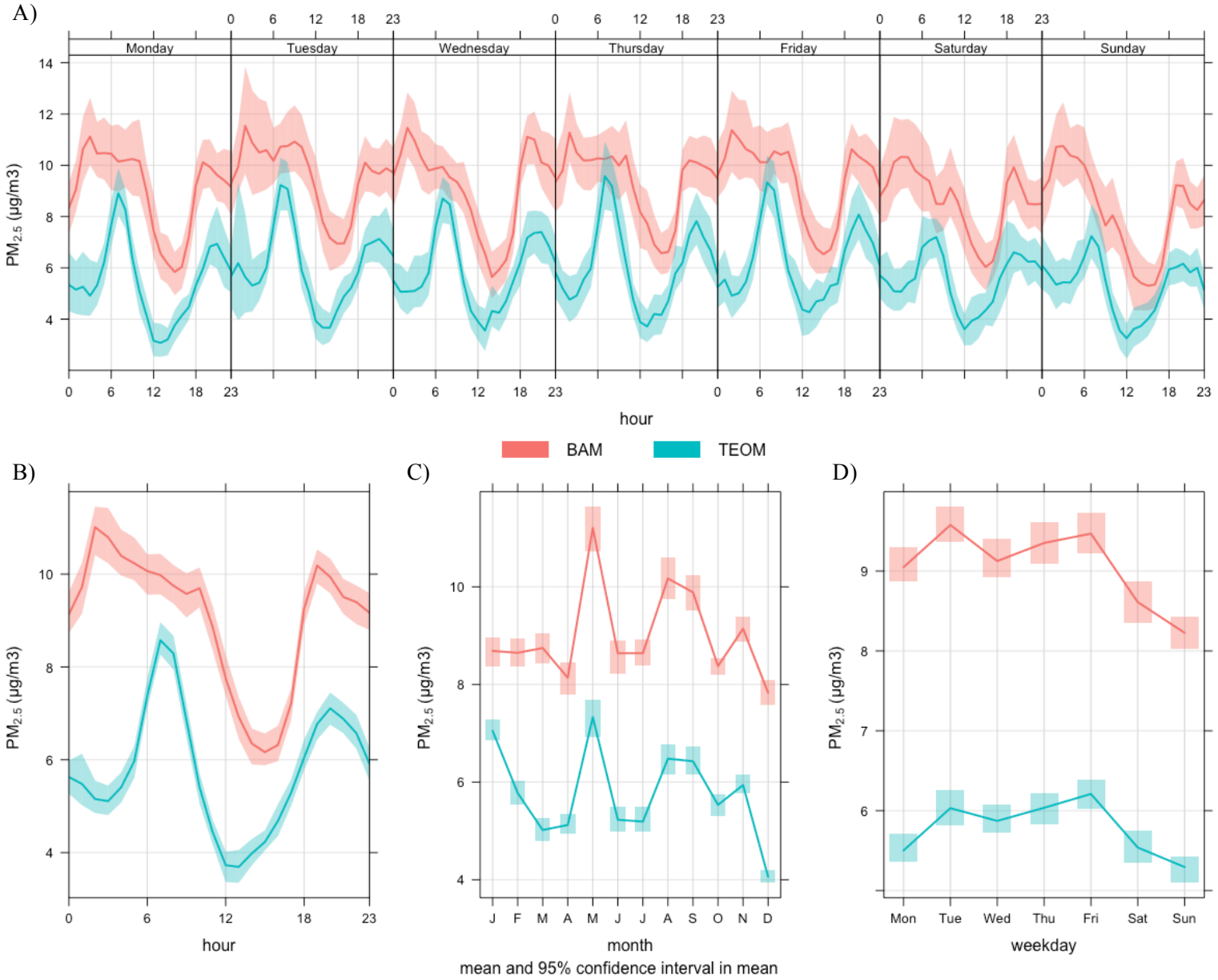


Figure 3- 5. Time variation plot of hourly data from collocated period, with the BAM in red and the TEOM shown in blue. The shading around the lines shows a 95% confidence interval. The plots show the A) hour-day, B) hour, C) monthly and D) daily averages.

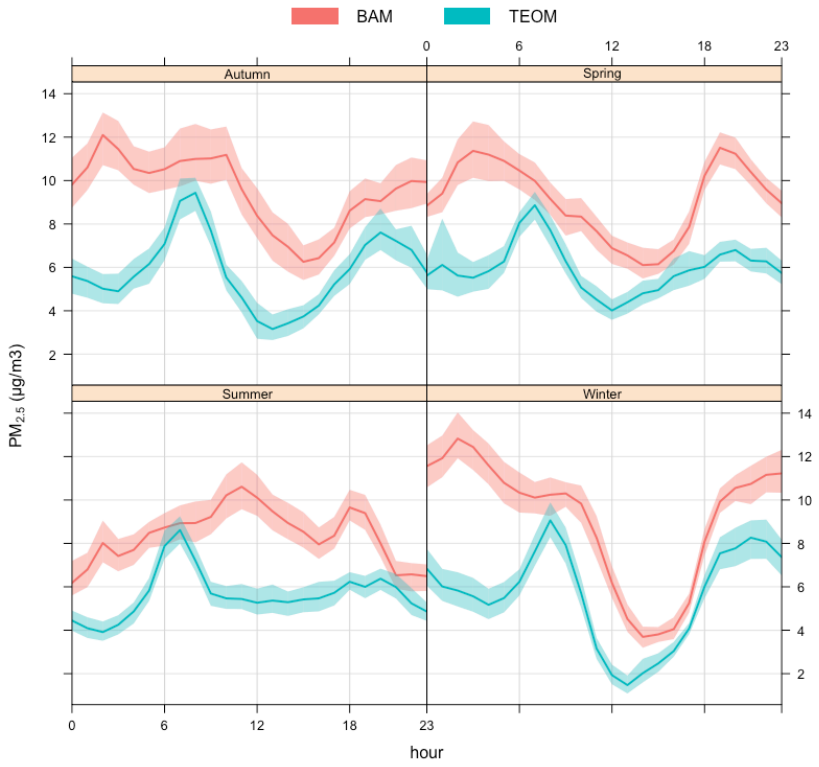


Figure 3- 6. Time variation for the collocated period; showing the hourly averages of the BAM (red) and TEOM (blue) readings divided by season for the study period.

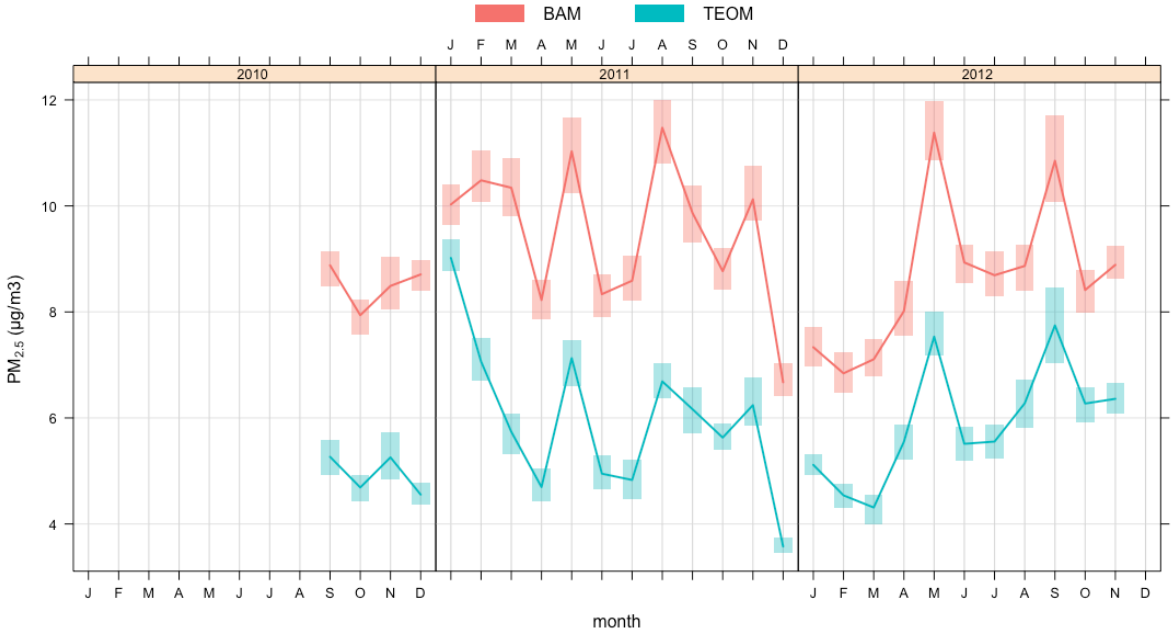


Figure 3- 7. Time variation for Chullora; showing the monthly average of the BAM (red) and TEOM (blue) readings, divided by year over the study period.

3.4 Correlation of $PM_{2.5}$ BAM with other variables

To assist in exploring the underlying structure of the data, the effect of meteorological conditions and other gaseous pollutants on $PM_{2.5}$ BAM was investigated. The descriptive statistics for variables over the collocated period is shown in Table 3- 1. Minimum values for the $PM_{2.5}$ TEOM and BAM and the PM_{10} TEOM were negative. In principle, such values are not possible. Maximum values for the BAM and TEOM occurred in spring offset by an hour of each other, on the 04/09/2012, at 01:00a.m. (TEOM) and 02:00a.m (BAM). This suggested that the BAM readings were lagged by an hour. The maximum value for the nephelometer also coincided with this date at 01:00a.m, suggesting a relationship between the nephelometer and the TEOM ($R^2 = 0.68$) and BAM ($R^2 = 0.41$). Corresponding results were observed for temperature, relative humidity, carbon monoxide, nitrogen monoxide, nitrogen oxides, nitrogen dioxide, sulfur dioxide, ozone, wind speed and wind direction in Table 3- 1.

3.5 Transforming data

The statistical techniques used in this project assume that the data has a normal distribution. This data possesses a strong asymmetry, with many tails in one end (Figure 3- 4). Hence, to improve the statistical properties of the data, the data was transformed by applying one mathematical function to all raw data values from a variable, and is further explained below.

Transforming for symmetry

Linear models rely on the assumption of a normal distribution. The distribution of the data in Figure 3- 4 shows us that the raw data from the TEOM and BAM do not have normal distributions. Hence, transformation methods were explored to identify which one produced the most normal distribution.

A log transformation is an appropriate (and a standard) transformation for atmospheric data (Bouis, 1999, Kahn, 1973) to reduce the skewness, and convert these distributions to a Gaussian distribution. However, log transformations are not defined for negative values. Here, it was assumed that the negative TEOM and BAM values were not actually true values, as in principle negative concentrations do not exist. However, these were still seen as valuable data which we did not want to get rid of. Therefore, the minimum value, plus 0.01 $\mu g/m^3$, of each the TEOM and BAM was added to each variable, shifting the data upwards.

After this, logarithmic and square root functions were trialed. Applying a logarithmic transformation resulted in the most normal distribution of TEOM and BAM values. The equation applied to transform the TEOM and BAM data is shown below:

$$Transformed\ PM_{2.5} = log\ (PM_{2.5}\mu g/m^3 + \ 2.51\ \mu g/m^3).$$

The transformed variables come from approximately the same distribution, displaying approximate normality (Figure 3- 9). Values less than zero appear to not fit as well to the normal distribution (Figure 3- 8 and Figure 3- 10), possibly as a result of smaller raw values having a larger measurement error. Once transformed, the majority of the TEOM and BAM data fits the normal density curve well (Figure 3- 10). The data with the highest density tends to exceed the normal distribution curve for both the BAM and TEOM, indicating a slight under dispersion; there is more data near the mean than a normal distribution should have.

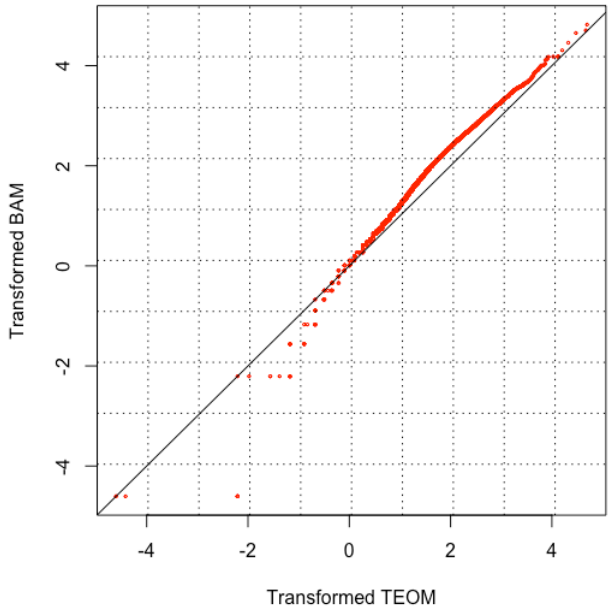


Figure 3- 8. Q-Q plot for transformed data, showing that the distributions are the same for the TEOM and the BAM.

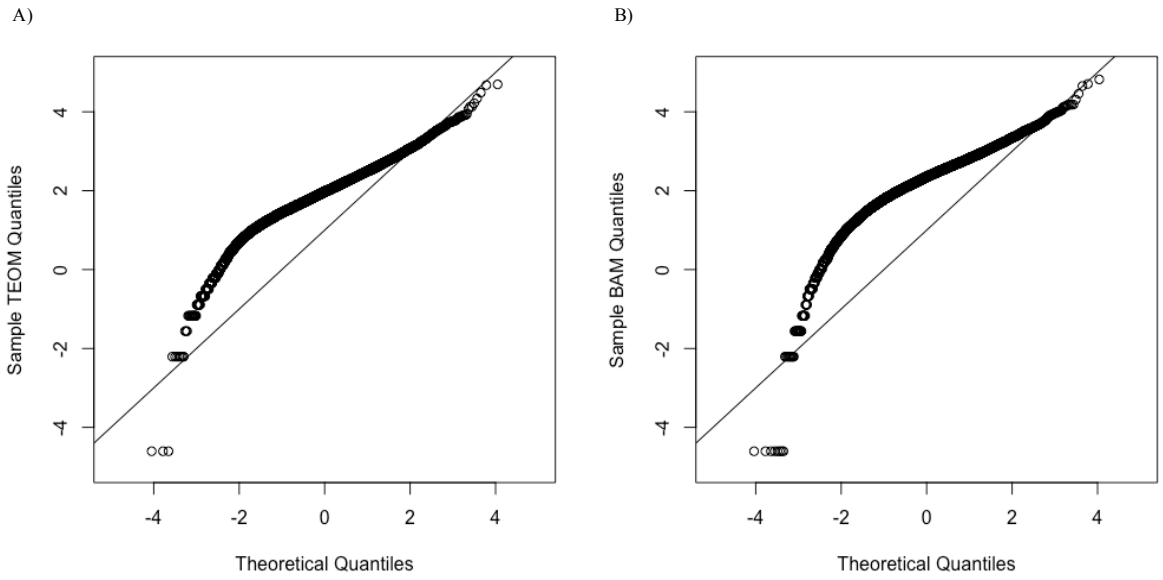


Figure 3- 9. Q-Q plots of theoretical vs actual quantiles, for A) TEOM and B) BAM.

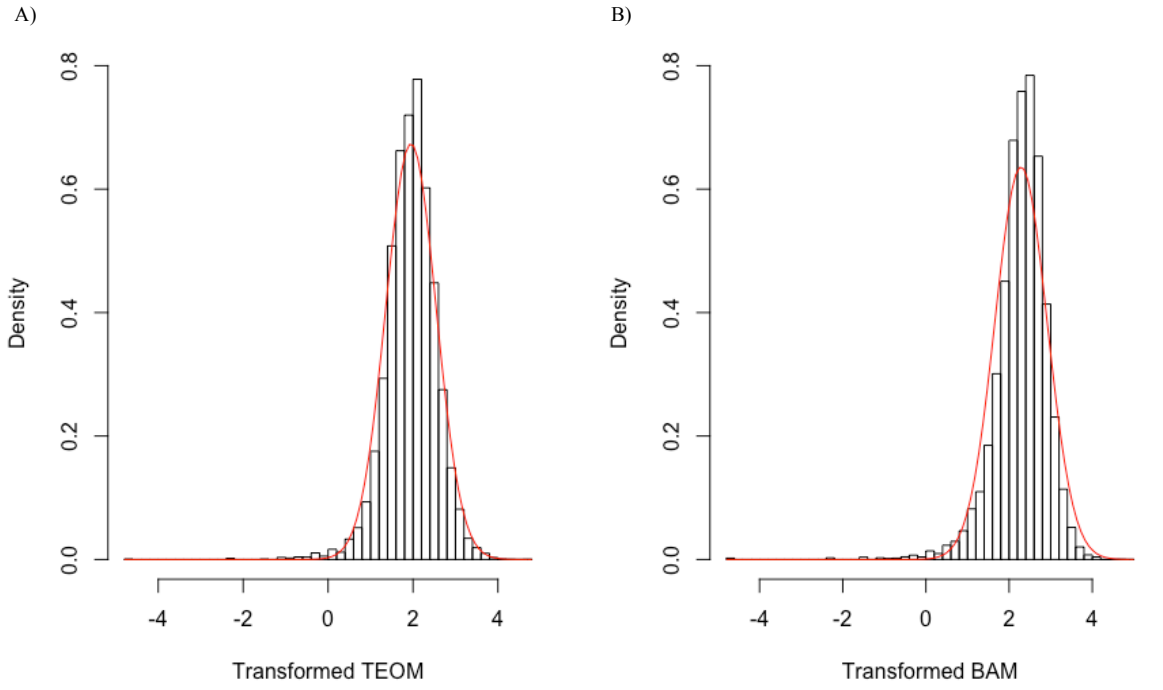


Figure 3- 10. Density histogram of transformed A) TEOM and B) BAM, with a normal density curve fitted, as shown in red. The mean and sample standard deviation were used to define this particular normal distribution curve.

Transforming for linearity

Our aim is to find linear relationships between air quality and meteorological variables, our predictors, and $PM_{2.5}$ BAM, for the prediction of $PM_{2.5}$ BAM. Our focus is on linear relationships as they are easily interpreted and departures from fit can be detected more easily. Transformation is a technique used that sometimes straightens relationships that were originally not linear.

To test if logarithmic transformation of the remaining gasses (CO, NO_x, NO, NO_2, SO_2, Ozone) improved their linearity with the $PM_{2.5}$ BAM, we used a method commonly used in the literature to transform gasses (Rosamond et al., 2012), that being:

$$x = \log(y_t + 1),$$

where *x* is the transformed value of the gas and *y* is the raw gas concentration at time *t*.

The value of 1 is added to the gas concentration, as when calculated, all raw untransformed negative values are ultimately excluded from the transformed data (negative log is undefined), and all raw untransformed zero values remain as zero values in the transformed data set ($log(0\ \mu g/m^3+1) = 0$).

For consistency, and to improve the distributional properties, PM_{10} underwent the same transformation process as the $PM_{2.5}$ readings. The most negative PM_{10} reading was 9.60 $\mu g/m^3$, therefore the following equation was applied to improve the straightness of the PM_{10}:

$$Transformed\ PM_{10} = log(PM_{10}\ \mu g/m^3 + 9.61\ \mu g/m^3).$$

NEPH, CO, NO_x NO_2 and NO all improved in their linearity with BAM when transformed. Ozone and SO_2 did not improve their linearity with transformed BAM when they themselves underwent transformation. Appendix 3 shows the linearity of all variables before and after their transformation.

3.6 Lagged variables

Models using lagged independent variables are named distributed lag (DL) models. Lagged independent or dependent variables are used in a model when it is evident that an independent or dependent variable that is lagged in time influences the dependent variable. Table 3- 4 displays the rho values between the dependent variable ($PM_{2.5}$ BAM) and lagged values of itself, along with three independent variables, including their lagged values. The results show that the lagged BAM values are critical in assisting with prediction of BAM, as demonstrated by a strong BAM lag 1 rho value (0.76). It is clear that in all cases, the agreement between the dependent variable (BAM) and the independent variables ($PM_{2.5}$ TEOM, PM_{10} TEOM, NEPH) improves when lagged values are used, at least up to lag 1. Values of lag 24 were investigated due to the possibility of readings at the time from the previous day being a useful predictor. In all cases, lag 24 was not favorable over the lag 0, 1 or 2 variables. The results from the correlations, show complex relationships not only at time 0, but also over time. All of these variables should ideally should be included in the model.

Table 3- 4. Relationship between $PM_{2.5}$ BAM and $PM_{2.5}$ BAM lagged values, along with three independent variables and their lagged values, based on hourly values. Lags are at hourly intervals and correlations are based on transformed variables.

Instrument	Independent variable	Correlation (Rho value)
$PM_{2.5}$ BAM	Lag 1	0.76
	Lag 2	0.56
	Lag 24	0.32
$PM_{2.5}$ TEOM	Lag 0	0.49
	Lag 1	0.54
	Lag 2	0.54
	Lag 24	0.29
PM_{10} TEOM	Lag 0	0.39
	Lag 1	0.42
	Lag 2	0.43
	Lag 24	0.19
NEPHELOMETER	Lag 0	0.56
	Lag 1	0.59
	Lag 2	0.57
	Lag 24	0.31

3.7 Stationarity

Many statistical forecasting methods are based on the assumption that the time series is approximately stationary. A stationary dataset is one whose values do not depend on the time at which the series is observed (Hyndman and Athanasopoulos, 2013). The statistical properties of the data, including mean, variance and autocorrelation are all constant over time. In some cases, stationarity can be achieved through transforming the data.

A Ljung-Box test was applied to test the stationarity of the time series using the transformed values previously calculated. The results from the Ljung-Box test suggest that the time series is non-stationary (*p-value* = 0.00).

Useful plots used to determine stationarity in a time series are autocorrelation function (ACF) and partial autocorrelation (PACF) plots. When a time series is stationary, the ACF will drop within the 95% limits immediately. Figure 3- 11 suggests that the time series is not stationary, as the ACF drops within the 95% limit relatively slowly.

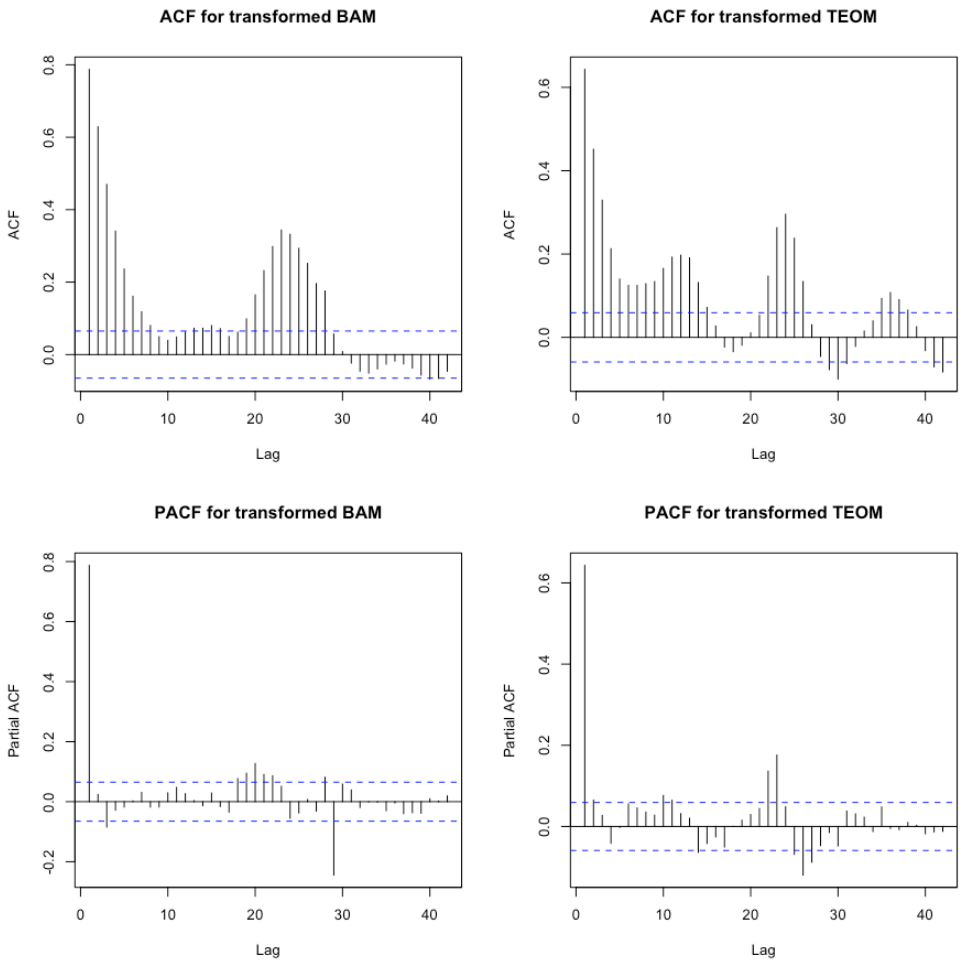


Figure 3- 11. ACF and PACF plots for the transformed TEOM and BAM values for the collocated period. The lags are at an hourly time scale.

3.8 Decisions and assumptions of model

Based on our exploratory data analyses, and the non-stationarity of the transformed data, we applied an autoregressive finite distributed lag (ARDL) model to address the main goal of this paper. Autoregressive (AR) models predict future behavior according to past behavior, and are a useful tool for forecasting when there is a correlation between values in a time series and those values that fall before and after them (lagged variables). The model is essentially a linear regression of the data, with the dependent variable directly related to the independent variable. An AR model differs from a simple linear regression as Y is dependent on X, and previous values of Y. As the agreement between the dependent and the independent values improved when values were lagged, and models that use lagged independent variables are named DL models, we will use a DL model in our prediction. When using an AR and DL model in combination, this is called an ARDL model.

A finite ARDL is appropriate for time series data, in which a regression equation is developed to predict values of a dependent variable based on current and lagged values of this explanatory variable. The starting point for the model takes the form of:

$$Y_t = a + w_0 x_t + w_1 x_{t-1} + w_2 x_{t-2} + \ldots + w_n x_{t-2} + error\ term,$$

where Y_t is the value at time t of the dependent variable y, a is the intercept term that is estimated, w_i is the lag weight on the value i periods previously of the explanatory variable x. The model is *finite* as there are a finite number of lag weights, signifying an assumption that there is a maximum lag beyond which values of the predictor variable do not influence the response variable. Lagged variables were added to the data frame in R and were offered in the pool of variables available for model construction.

An ARDL model holds the same assumptions of a linear regression model. Therefore, there are five key assumptions to be met, they are 1) a linear relationship, 2) little multicollinearity, 3) little or no auto-correlation, 4) normality of the residuals and 5) homoscedasticity. Each of these assumptions is discussed below.

Firstly, the relationship between the independent and dependent variables needs to be linear. The linearity is best expressed with scatter plots, as previously discussed in *3.5 Transforming data* and shown in Figure AP3-1 and Figure AP3-2 in Appendix 3. Variables that do not meet this assumption include Ozone and SO_2.

Secondly, linear regression assumes there is little or no multicollinearity present in the data. Multicollinearity occurs when independent variables are too highly correlated with each other, and can cause an overfitting of the regression analysis model and instability. It is important to test and remove multicollinearity as it can cause imprecise estimates of coefficient values, and impact the out-of-sample predictions. Multicollinearity is tested through a correlation matrix, and is shown in Table AP3-1 of Appendix 3. Variables with a correlation coefficient of ≥ 0.8 were deemed to be collinear. Variables excluded from the analyses due to collinearity with variables already included in the model include NEPH, NEPH lag 2 and NO_x. Given the high degree of multicollinearity between NEPH, NEPH lag1 and NEPH lag 2, NEPH and NEPH lag 2 were omitted from the pool of variables available for selection. The reason for keeping NEPH lag 1 over the other variables was because it expressed the highest correlation with the BAM value (rho = 0.59, as opposed to rho = 0.56 and rho = 0.57 with the NEPH and NEPH lag 2 respectively). NO_x was omitted due to its high correlation with CO (rho = 0.79).

Additionally, there must be little or no auto-correlation in the residuals. Autocorrelation occurs when the residuals are not independent from each other, which will be examined once the model is developed by observing the ACF and PACF plots.

The residuals of the model need to be normal, for the purpose of meeting the assumption of normality for the statistical tests that will be performed once the model has been developed (e.g. prediction and confidence intervals will be imprecise if the model is not normal). Normality can be checked once the model has been constructed.

Lastly, an ARDL model assumes homoscedasticity. This means that the residuals are equal across the regression line. This too will be discussed in Chapter 4, when the model has been made and is under evaluation.

3.9 Summary

Based on our exploratory data analyses, an auto-regressive finite distributed lag model is appropriate for this particular application. The majority of the assumptions of the model are now met, with the remaining assumptions needing to be tested once the model has been constructed.

Chapter 4: Model building and evaluation

4.1 Overview

Results from our exploratory data analyses reveal the appropriateness of applying an ARDL model to predict what the $PM_{2.5}$ values would have been, in the years before the BAM instrument was installed. Now we have thoroughly explored the data, we will build the predictive model and evaluate its effectiveness. The aim of this chapter is to produce the following:

a) A list of outliers
b) A good-fitting, parsimonious model
c) Estimates for parameters
d) Uncertainties for estimates
e) A ranked list of important factors
f) A sense of robustness of conclusions

4.2 Data preparation

4.2.1 Dealing with missing values

The modelling of air quality trends largely rests on statistical analysis of data collected at monitoring stations. However, it is common that not all scheduled measurements are made. The reasons for missing data in a set may include machine failure, routine maintenance, human error or other factors. It is acknowledged that incomplete datasets may produce results that vary from those that would have been acquired from a complete dataset (Hawthorne and Elliott, 2005), with data-base users often obliged to complete the data sets themselves. Imputation is a common method used to determine a value for missing values in a dataset. However, in this analysis we chose to leave missing data as NA's, as the development and testing of a model for imputing the missing values was beyond the scope of this project (the development of a statistical multiple imputation model that accurately reflects the trend, seasonal cycle, and joint error structure of multiple atmospheric gases is time-consuming due to the extensive analysis needed to evaluate the most appropriate imputation technique with the assumptions of the imputation needing to be checked), and the use of a simple imputation technique that did not accurately reflect the joint distribution of the variables at any given timepoint may have biased our results.

4.2.2 Outlier detection and removal

Outliers are data points that deviate significantly from others, and are a challenge to properly deal with in science research. The different methods of defining, identifying, and handling outliers can significantly change study conclusions (Aguinis et al., 2013). As emphasized by Cortina (2002), "caution also must be used because, in most cases, deletion [of outliers] helps us to support our hypothesis" (p.359). Removing outlier values can be problematic, having the capacity to cause favorable results that produce a model with a better fit. However, it should also be mentioned that outliers present in data can have such a strong influence on the data that they bias the fit estimators, predictors and accuracy of the model. Ultimately, it becomes a tradeoff, and is left to the researcher to decide on the appropriateness of removing outliers.

Visual inspection of the TEOM and BAM measurements reveal cases where the values were quite different, to such an extent that one would expect they represent erroneous data. However, given that the purpose of this study is to identify any evidence of differences in responses by the $PM_{2.5}$ TEOM and BAM, standard air quality data editing practice was limited, as it might remove data that reflects real biases in each of the samplers. Hence, a conservative approach to data consistency was applied, allowing the inclusion of data displaying significant levels of inconsistency.

The data was visually inspected to identify outliers, via a scatter plot (Figure 3- 1A) in combination with a plot of residuals against leverages (Figure 4- 1). From Figure 3- 1 A, it seems that the data point furthest to the right on the *x*-axis is an outlier (identified as TEOM = 170.9 $\mu g/m^3$, BAM = 45.1 $\mu g/m^3$ on 04/09/2012 at 01:00a.m). The point appears to not follow the same trend when compared to the rest of the data.

Figure 4- 1 assists with identifying influential points that influence the regression line, by showing Cook's distance, indicated by the red dotted line. Cook's distance measures the effect of removing a certain observation, with the larger the distance indicating a more influential observation. Observations outside of the red dashed line are influential to the regression results. Data from row 17, 577 was identified as an influential observation. This is the same data point identified from the visual inspection. The TEOM value was removed from the dataset and assigned an NA value, due to concerns that it might influence the predictive ability of the model given its extreme value.

4.2.3 Lagging

Due to our exploratory analysis revealing the significance of lagged variables, it is important to include lagged variables in the pool of covariates available for model construction. This lag is added as a new variable to the dataset, and is called a lag-response, in addition to the standard exposure-response relationship. Lag-response variables of 1hr, 2hrs and 24hrs for the $PM_{2.5}$ TEOM, nephelometer and PM_{10} measurements were made available as variables for selection in the model.

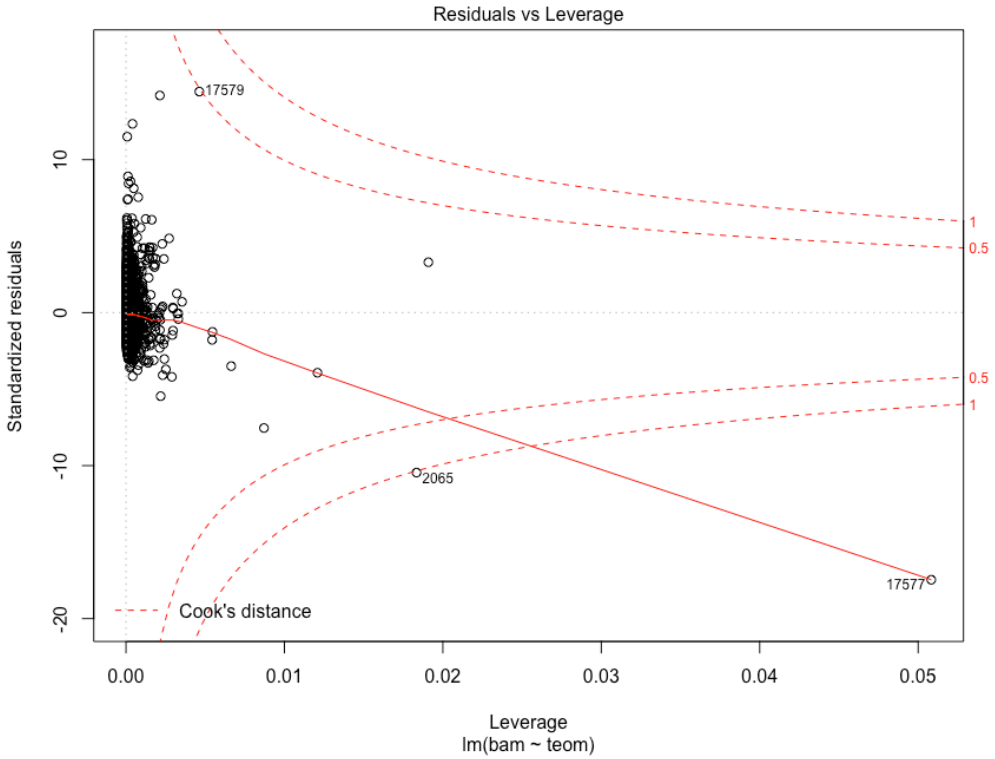


Figure 4- 1. A plot of residuals against leverages, along with Cook's distance.

4.2.4 Breaking up monthly and hourly data into blocks

Data blocking was performed to satisfy the aim of constructing a parsimonious model. That is, a model that accomplishes a good level of prediction while using as few variables as possible, without sacrificing rigor. It would be inappropriate to input 24 hourly values and 12 monthly values into the model as this would lead to such a large number of input variables. Hence, we blocked the hourly and monthly data based on their significance levels. Hourly values were segmented into block *a*: 11:00p.m. – 2:00a.m., *b*: 3:00a.m. -7:00a.m., *c*: 8:00a.m. – 3:00p.m. and *d*: 4:00p.m.-10:00p.m. These statistical blocks of hours appear to be reasonably physical since they seem to follow the diurnal gas cycle, with low values

overnight, a fall and rise of values during the day, and higher values during morning and afternoon peak periods. Monthly data was blocked into block *a*: November to March, and *b*: April to October. Again, the monthly statistical blocks follow reasonably well with the physical cause. November through to March have a fairly constant average $PM_{2.5}$ readings, with the months of April through to October possessing more variation. Reasons for the statistical cut-off points for these blocks is explained in Appendix 4.

4.3 Variable selection and model construction

As there are many independent variables available for selection (Table 4- 1), a strategy needs to be employed to select the best predictors to use in the regression model. Therefore, four measures of predictive accuracy were incorporated to determine the best model. They are Adjusted R^2 ($\bar{R}^2$), Cross Validation (CV), Akaike's Information Criterion (AIC) and Schwarz Bayesian Information Criterion (BIC), and are defined in Appendix 5.

Table 4- 1. Table of variables available to be used in the predictive model.

Variable Name	What is it?
teoml	log(teom μg/m^3 + 2.51)
teoml.l1	Lag1(teoml)
teoml.l2	Lag2(teoml)
teoml.l24	Lag24(teoml)
nephl.l1	Lag1(log(neph))
nephl.l24	Lag24(log(neph))
pm10l	log(pm10 μg/m^3 + 9.61)
pm10l.l1	Lag1(pm10l)
pm10l.l2	Lag2(pm10l)
pm10l.l24	Lag24(pm10l)
temp	Temperature
rh	Relative Humidity
lco	Log(COppm + 1)
lno2	Log(NO_2ppm + 1)
lno	Log(NOppm + 1)
ws	Wind speed
wdir	Wind direction
mthbk	Monthly data broken into blocks based on significance levels.
hrbk	Hourly data broken into blocks based on significance levels.

CV is more accurate for smaller values of n, while $\bar{R}^2$ has a tendency to select too many variables, and BIC has a tendency to select too few variables. Therefore, priority was given to favorable AIC values first, then other measures were assessed as a secondary evaluation.

It would be unwise to fit all potential regression models (given there are 19 covariates available) and assess their measures of predictability, as there are more than 250,000 possible models. Therefore, two methods for variables selection were carried out, and their AIC, BIC, CV and adjusted R^2 were examined to determine the model with the best predictive ability. These two methods were manual f-test backwards selection and manual f-test forward selection. The specifics of these methods is explained in Appendix 6.The results of the methods for variable selection is shown in Table 4- 2.

Both methods of variable selection ended up selecting the same variables for the final hourly model . The closeness of the two models, in terms of their selected variables and measures of predictive ability, highlights the robustness of the variable selection method. The model summary is shown in Figure 4- 2. Note that this model is built on 16,711 complete observations, to assist in producing the best possible model.

Table 4- 2. Results from the model produced from the manual f-test forward and back selection – the same variables were chosen for both methods. Measures of predictive ability are shown by the CV, AIC, BIC and $\bar{R}^2$.

CV	**AIC**	**BIC**	$\bar{R}^2$
0.222	-23828.630	-23677.880	0.430

```
Call:
lm(formula = baml ~ temp + rh + hrbk + mthbk + teoml + teoml.l1 +
    teoml.l2 + teoml.l24 + nephl.l1 + pm10l.l1 + pm10l.l2 + lco +
    lno2 + lno + ws, data = mds)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9576 -0.1576  0.0480  0.2330  1.8599

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.5424929  0.0885141  28.724  < 2e-16 ***
temp        -0.0146844  0.0011929 -12.310  < 2e-16 ***
rh          -0.0054805  0.0003288 -16.667  < 2e-16 ***
hrbkb        0.1060767  0.0140436   7.553 4.47e-14 ***
hrbkc       -0.0841233  0.0142143  -5.918 3.32e-09 ***
hrbkd        0.0593728  0.0135268   4.389 1.14e-05 ***
mthbkb      -0.1041434  0.0118872  -8.761  < 2e-16 ***
teoml        0.0864400  0.0113426   7.621 2.66e-14 ***
teoml.l1     0.0537523  0.0165574   3.246  0.00117 **
teoml.l2     0.1343761  0.0144373   9.308  < 2e-16 ***
teoml.l24    0.0600673  0.0071617   8.387  < 2e-16 ***
nephl.l1     0.3592783  0.0105505  34.053  < 2e-16 ***
pm10l.l1    -0.1537244  0.0244864  -6.278 3.52e-10 ***
pm10l.l2     0.1518855  0.0233496   6.505 8.01e-11 ***
lco          0.5091722  0.0522786   9.740  < 2e-16 ***
lno2         0.0539037  0.0112575   4.788 1.70e-06 ***
lno         -0.0213109  0.0054326  -3.923 8.79e-05 ***
ws           0.0443205  0.0040665  10.899  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4714 on 15836 degrees of freedom
  (3797 observations deleted due to missingness)
Multiple R-squared:  0.4306,  Adjusted R-squared:   0.43
F-statistic: 704.5 on 17 and 15836 DF,  p-value: < 2.2e-16
```

Figure 4- 2. Model summary for predicting BAM hourly values.

4.4 Examining residuals

Exploratory data analyses heavily relies on the examination of residuals, as they assist with understanding the data and models. A data point can be described as: *data = fit + residual*. The fit captures the major trend of the data, with examination of the residuals enabling a more detailed understanding of the fit. The goal is to fit as much of the pattern in the data into the fitting technique as possible, while not fitting any noise in the data.

A good model contains no or little patterns in the residuals. ACF and PACF plots are useful to examine the residuals. The residuals of the model are plotted in Figure 4- 3. Ideally, the lags for the ACF and should drop within the 95% limits immediately, as a linear model assumes that the variance of the residuals are constant (i.e. independent) over the values of the response variable, indicating that the residuals are not autocorrelated. For our model, we

can see that there is still some autocorrelation between the residuals. This could be due to non-stationarity that this simple type of modelling is unable to remove.

Sometimes autocorrelation in the residuals can be solved by adding differenced variables; the change in a value from on period to the next. For example, if Y_t is the value of the time series Y at time t, then a first difference of Y at time t is equal to $Y_t - Y_{t-1}$. Second differencing involves differencing the differences, and is equal to $[(Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$. The TEOM, NEPH and PM_{10} data were differenced by first-order and second-order and were added to the pool of variables. When a new predictive model was made with these included as predictor variables, they did not reduce the autocorrelation expressed through the ACF and PACF plots, or improve the AIC, BIC, CV or $\overline{R}^2$ values.

Therefore, we can infer that there is information left in the residuals that should be included when computing the forecast (Hyndman and Athanasopoulos, 2013). Such information may include particle composition data or other air quality parameters. However, we have not considered data of this nature in our research.

Interestingly, if we add BAM lagged values to a new model, it significantly reduces the autocorrelation in the residuals (Figure 4- 4). The output for this model, with variables selected, parameter estimates and *p-values* is shown in Appendix 7. The R^2 for the predictive model rises from 0.43 to 0.67 when BAM lagged values are added to the model. Plus, the measure of predictive ability significantly improves (Table 4- 3 compared to Table 4- 2). However, we do not proceed with this model as we do not have access to such data at this point. Such BAM data could be sourced, and then adjusted, from another site that is in close proximity to Chullora. We are missing data from our perfect model, because the perfect model would include lagged values of BAM.

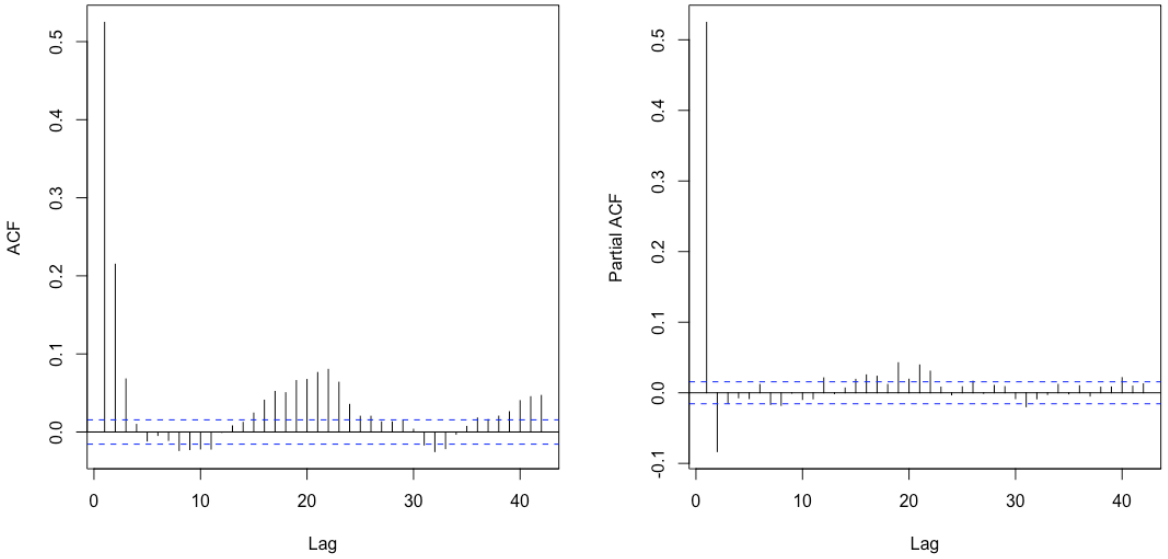


Figure 4- 3. ACF and PACF plots of final model used for prediction of BAM hourly values.

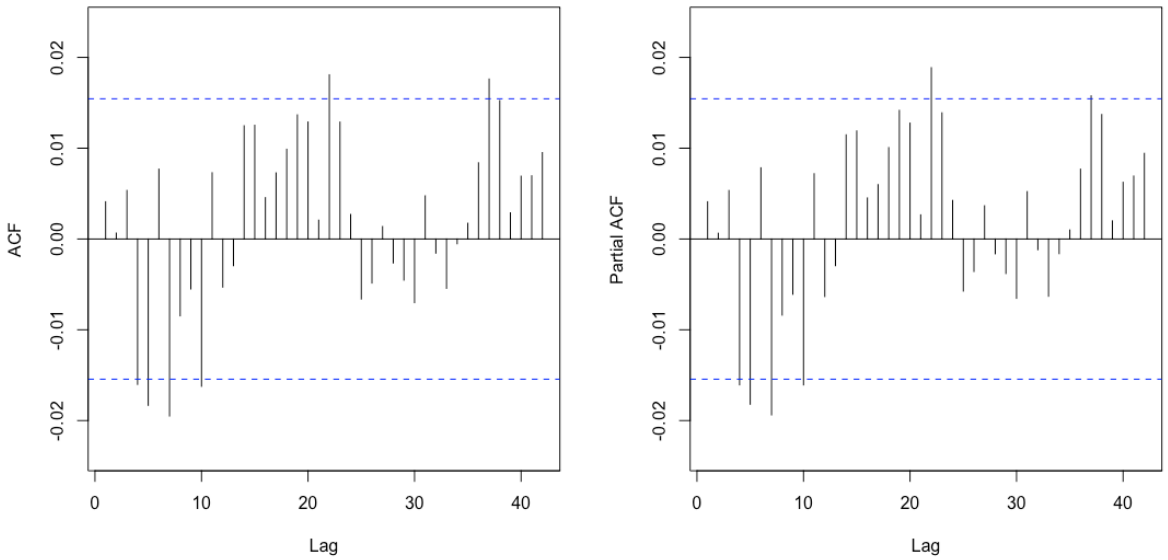


Figure 4- 4. ACF and PACF plots of a second model, that includes BAM lagged values as predictor variables.

Table 4- 3. Results from predictive model that include BAM lagged variables in its predictor variables. Measures of predictive ability are shown by the AIC, BIC, CV and $\bar{R}^2$.

CV	AIC	BIC	$\bar{R}^2$
0.118	-34428.600	-34274.860	0.669

The cross correlation function (CCF) plots reveal the suitability the variables to be used in a model. They measure the correlation between lagged values of two (or more) variables. The *x*-axis indicates the lag and the *y*-axis indicates the correlation. Cross correlation functions are used here to identify correlations between the residuals of the model and lagged values of covariates that have been included in the model (Figure 4- 5 and Figure 4- 6). These are checked to ensure that non-stationarity is not included in the model from the time-series structure of the predictors. There are exceedances of the 95% limit sometimes for some covariates, but most of the time the variable lie within this limit (Figure 4- 5 and Figure 4- 6). These exceedances suggest that there is some non-stationarity included in the model as a result of the predictors, but not enough to cause concern.

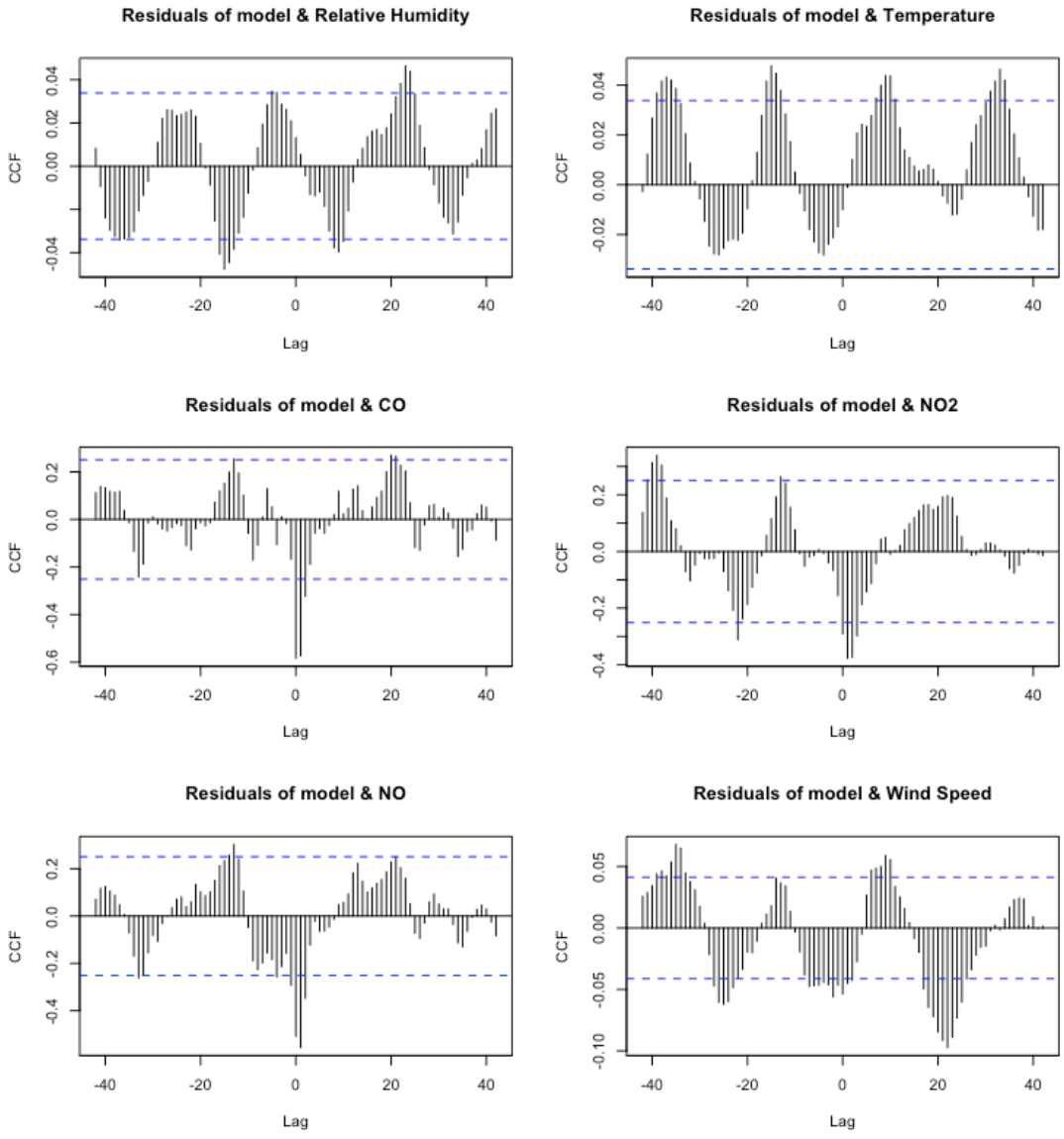


Figure 4- 5. CCF plots of meteorological and air quality variables against prediction model.

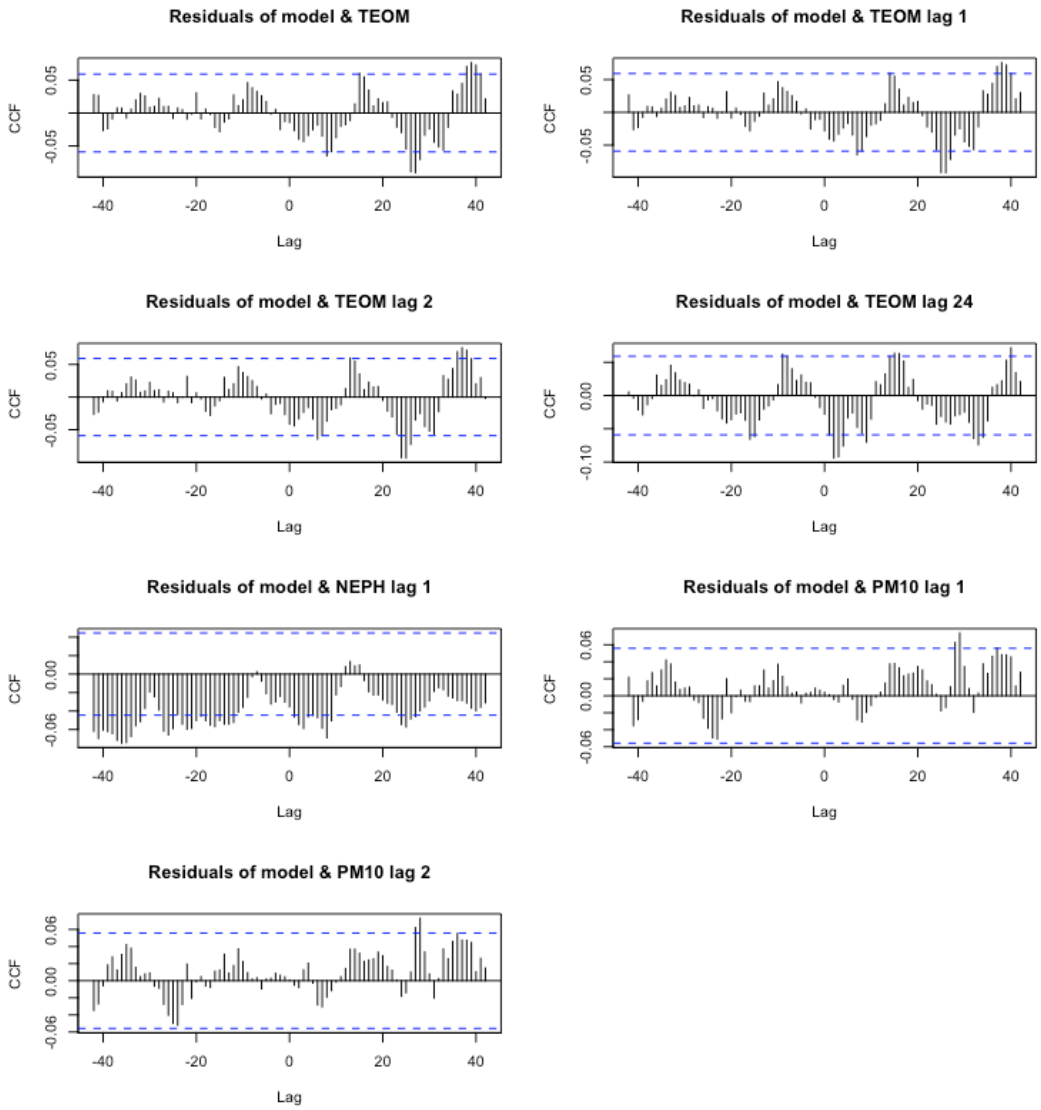


Figure 4- 6. CCF plots of PM_{10}, $PM_{2.5}$ and nephelometer predictor variables against prediction model.

4.5 Testing remaining assumptions of model

There are two remaining assumptions that need to be checked to ensure all assumptions of the model are met. They are homoscedasticity and normality of the residuals.

Testing for homoscedasticity

One of the key assumptions of our ARDL model was that the model is homoscedastic. This means that the variance around the regression line is the same for all values of the independent variable. The residuals occur randomly around the zero line (Figure 4- 7), indicating the suitability of assuming a linear relationship. The residuals roughly form a horizontal band around the zero line (Figure 4- 7), suggesting the variances of the error terms

are equal. And lastly, no one residual stands out from the pattern of residuals (Figure 4- 7), suggesting there are no outliers in the data set. All of these indicate homoscedasticity of the model.

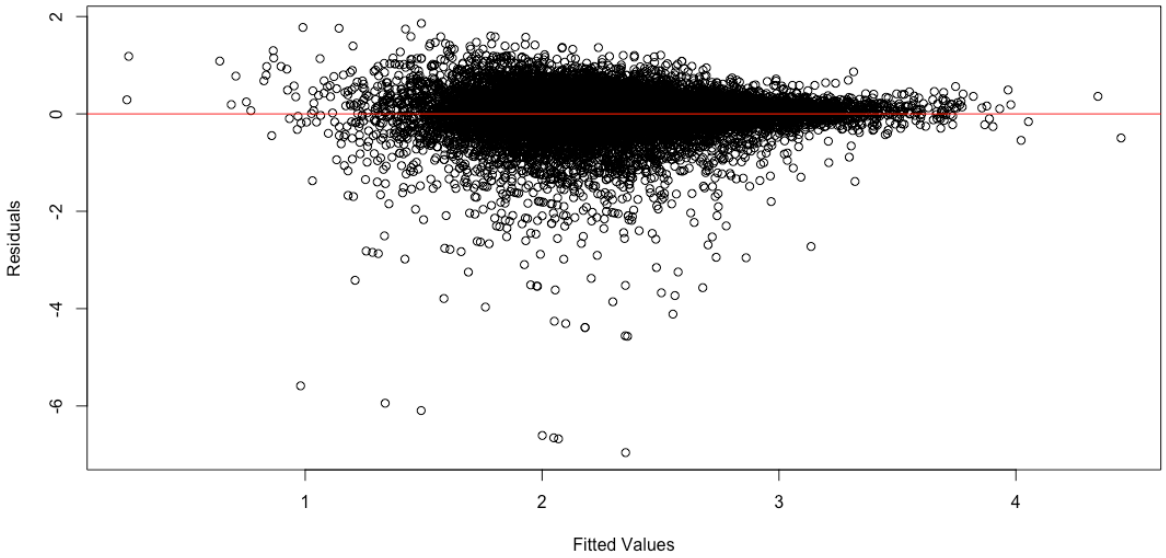


Figure 4- 7. Plot of residuals vs fitted values for the final hourly model.

Testing for normality of the residuals

The final assumption of the ARDL model was tested, that being that the residuals have a normal distribution. A Q-Q plot of the studentized residuals from a linear model against the theoretical quantiles of a comparison distribution indicate strong tailing to the left (Figure 4- 8 A). The histogram of the residuals suggests the distribution is not bell shaped, but negatively skewed (Figure 4- 8 B). Both of these suggest non-normality of the residuals, however, such deviation is not a concern for large datasets (Lumley et al., 2002). Given the large amount of data, the violation of the assumption is not so important. It will not have a substantial impact on the conclusion of the model.

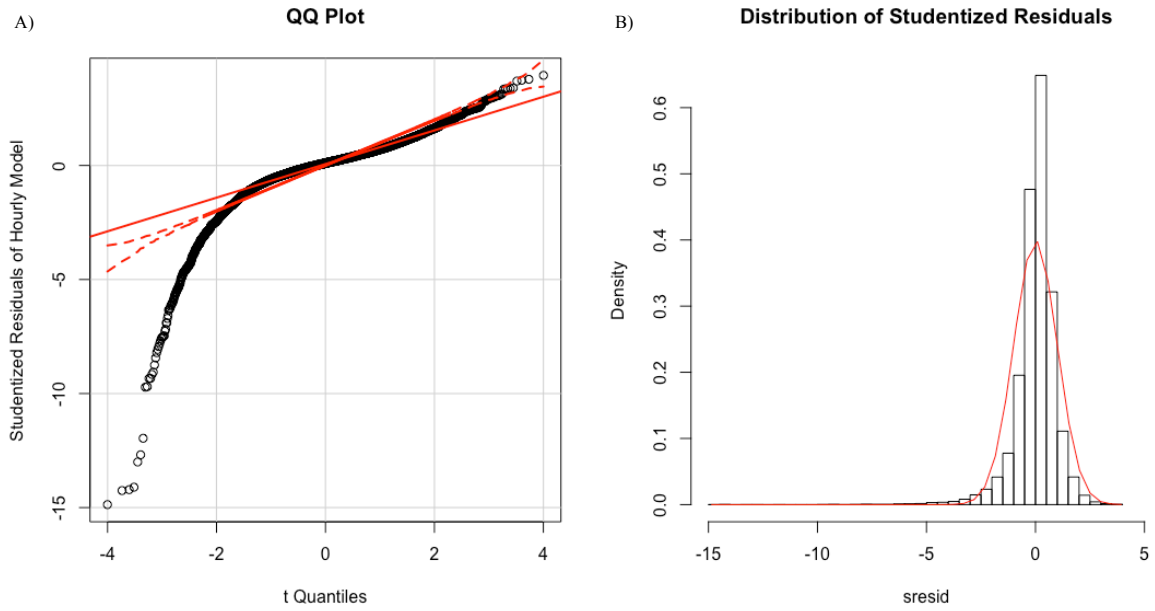


Figure 4- 8. A) Q-Q plot of studentized residuals from the daily model against theoretical quantiles. B) Histogram of studentized residuals. The red line indicates a normal distribution, as calculated from the minimum and maximum studentized residuals.

4.6 Measures of accuracy

Evaluating the models performance is important to establish its credibility in simulating the actual $PM_{2.5}$ readings. Confidence intervals are relevant for parameter estimates and mean estimates. They show how precisely we know the estimate. They indicate the mean response for a particular value of *x*. The 95% confidence interval of the mean predicted transformed BAM values is between 2.25 and 2.31 (this equates to 6.97 µg/m^3 and 7.58 µg/m^3). Therefore, there is a 95% probability that the interval we obtained contains the true value of BAM $PM_{2.5}$ at the specified model settings. The confidence intervals for the parameter estimates indicate the likely range of the true, unknown parameter, reflecting the amount of random error in the sample. These are shown in Table 4-4.

Prediction intervals are relevant for predicting observations, indicating what value will the response be assuming a particular value of *x*. The 95% prediction interval of the mean predicted transformed BAM values is between 1.35 and 3.20 (this equates to 1.37 µg/m^3 and 22.12 µg/m^3).

The residual standard error of a model is also a useful tool to evaluate how well the model fits the data. The standard error for this model is 0.4714 on 15287 degrees of freedom.

Table 4- 4. Parameter estimates and confidence intervals for hourly predictive model.

		Confidence interval	
Parameter	**Estimate**	**2.50%**	**97.50%**
(Intercept)	2.542	2.369	2.716
temp	-0.015	-0.017	-0.012
rh	-0.005	-0.006	-0.005
hrbkb	0.106	0.079	0.134
hrbkc	-0.084	-0.112	-0.056
hrbkd	0.059	0.033	0.086
mthbkb	-0.104	-0.127	-0.081
teoml	0.086	0.064	0.109
teoml.l1	0.054	0.021	0.086
teoml.l2	0.134	0.106	0.163
teoml.l24	0.06	0.046	0.074
nephl.l1	0.359	0.339	0.380
pm10l.l1	-0.154	-0.202	-0.106
pm10l.l2	0.152	0.106	0.198
lco	0.509	0.407	0.612
lno2	0.054	0.032	0.076
lno	-0.021	-0.032	-0.011
ws	0.044	0.036	0.052

4.7 Model validation and evaluation

Evaluating a models performance is a crucial step so accurate conclusions can be drawn from the research. One way to evaluate the forecast accuracy of the model is to perform forecasts on an independent test set of data. Since the time series data exhibits strong autocorrelation, and lagged variables were included in the model, conventional 10-fold cross validation was not used. Instead, time-series cross validation was implemented, enabling multiple rounds of forecasts to obtain more reliable forecast accuracy measures (Arlot and Celisse, 2010), whilst preserving independent observations to test the model on (Hyndman and Koehler, 2014). This method uses many training sets of data, each one containing one more observation than the preceding one. Hourly one-step cross validation process was used (Figure 4- 9).

For this method, the first M observations are used to train the model. Then, the covariates for observation $M + 1$, including lagged values of observation $M + 1$, are used to obtain a prediction for the observation $M + 1$. Next, we used the $M + 1$ observation to re-train the model (1 year plus one hour) and obtain a prediction for $M + 2$. Then, we used the first $M + 2$ observations to re-train the model (1 year plus two hours) and obtain a prediction for $M + 3$. This procedure was applied until the first $N – 1$ observations were used to train the model and a prediction was obtained for N.

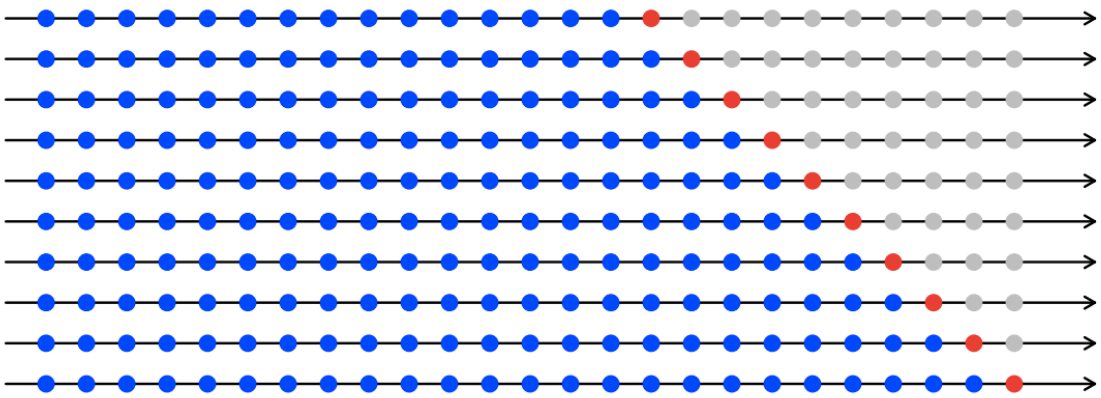


Figure 4- 9. Time-series cross validation based on one-step forecast. The blue points indicate the training set, the red points indicate the test sets and the grey points are ignored. Image sourced from Hyndman and Koehler (2014).

In this case, we use one year's worth of data to train the model (*M;* from 02/09/2010 to 02/09/2011), as variations from all hours, days, months and seasons were captured. The model was then applied on just over a year's worth of data (from 03/09/2011 to 29/11/2012), and the results were examined.

The `modStats` function, from the *Openair* library, was applied to statistically evaluate the model. The statistical output includes: fraction of predictions within a factor of two (FAC2), mean bias (MB), mean gross error (MGE), normalized mean bias (NMB), normalized mean gross error (NMGE), root mean squared error (RMSE), the Pearson correlation coefficient (r), the coefficient of efficiency (COE) and the index of agreement (IOA). These parameters are defined in Appendix 8. The summary statistics of these predictions are shown in Table 4- 5. The model was applied on 8,767 (n) observation of data. Table 4- 5 shows that the large majority of predictions are within a factor of two – ranging from 0.69 in summer to 0.90 in winter. A perfect model will have an FAC2 of 1.0. The MB has a negative bias, therefore underestimation of modelled values in all seasons. The underestimation is greatest in autumn (1.01 $\mu g/m^3$) and least in winter (-0.21 $\mu g/m^3$). The MGE shows the most in summer (3.45 $\mu g/m^3$) and the least in winter (2.03 $\mu g/m^3$). As MGE is calculated using absolute values, we can conclude that the summer modelled values possess the greatest spread from the observed values. Also the correlation coefficient is considerably lower in summer (r = 0.39) compared to spring, autumn and winter (r = 0.83, 0.82 and 0.91 respectively). The RMSE is greatest in summer (4.76) and lowest in winter (2.71). The COE indicates that the model is superior winter (COE=0.58), compared to the other seasons (spring COE= 0.41, autumn COE = 0.41, summer COE = 0.11), as models with

a COE closer to one performs better. The model performs so poorly in summer with a COE of 0.11, that it can be said that the model only predicts the observed values slightly better than it would using the observed mean. Lastly, the IOA is highest in winter (0.79) and the lowest in summer (0.56). Models with an IOA approaching + 1 represent a better model performance. Seasonally, we conclude that the models predictive ability is very poor in summer as demonstrated by the statistical output in Table 4- 5, and discussed above. The predictions for spring and autumn perform pretty similar, with the model possessing its greatest predictive ability in winter.

Exploring the statistical output for the models predictive ability overall, we conclude that the models performance is only satisfactory. The model underestimates observed $PM_{2.5}$ values, as demonstrated by the MB of -0.43 $\mu g/m^3$. An r of 0.80 suggests a strong correlation between the observed and modelled values, but it is crucial to consider that correlation alone should not be used to assess agreement between the two (Mukaka, 2012, Schweizer et al., 2016). The COE is 0.41, and IOA is 0.70, indicating there is still a lot of room for improvement of the hourly model.

Table 4- 5. Common numerical model evaluation statistics, based on predicted value from time series cross validation.

season	n	FAC2	MB	MGE	NMB	NMGE	RMSE	r	COE	IOA
spring (SON)	3444	0.850	-0.288	2.812	-0.030	0.298	4.019	0.828	0.413	0.707
summer (DJF)	1719	0.686	-0.393	3.445	-0.056	0.494	4.755	0.388	0.110	0.555
autumn (MAM)	1685	0.776	-1.008	3.058	-0.113	0.343	4.110	0.821	0.414	0.707
winter (JJA)	1919	0.896	-0.206	2.027	-0.023	0.226	2.708	0.905	0.575	0.788
All data	8767	0.814	-0.429	2.811	-0.049	0.321	3.953	0.804	0.407	0.703

With the statistical analysis above providing a lot of important information on the models performance, it is limited to numerical outputs. A much richer source of information on model performance is explored below, using graphical representation of the data, to assist in answering why the model is not performing well.

A scatter plot of the hourly actual BAM readings and the predicted BAM readings is shown in Figure 4- 10, from the predictions made from the time-series cross validation. There is fair agreement between the BAM and predicted BAM values ($R^2 = 0.44$). There is still a lot of scatter either side of the regression line. The standard error for the intercept and slope coefficient are 0.03 and 0.01 respectively.

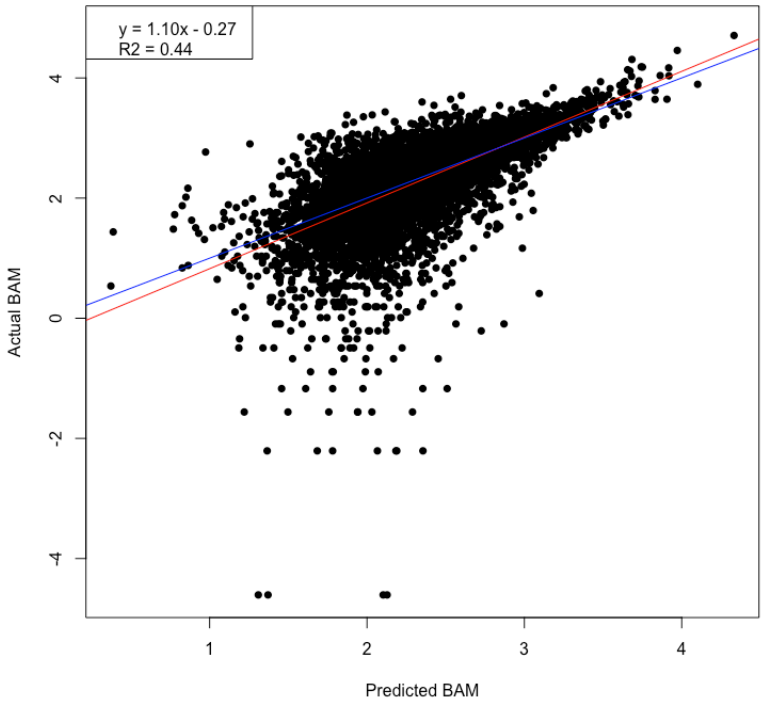


Figure 4- 10. A scatter plot for the transformed actual BAM values and transformed predicted BAM values, based on hourly values, for values produced from the time-series cross validation period. The ordinary least squares regression line is displayed in red, a 1:1 line is shown in blue, and the coefficients are also presented.

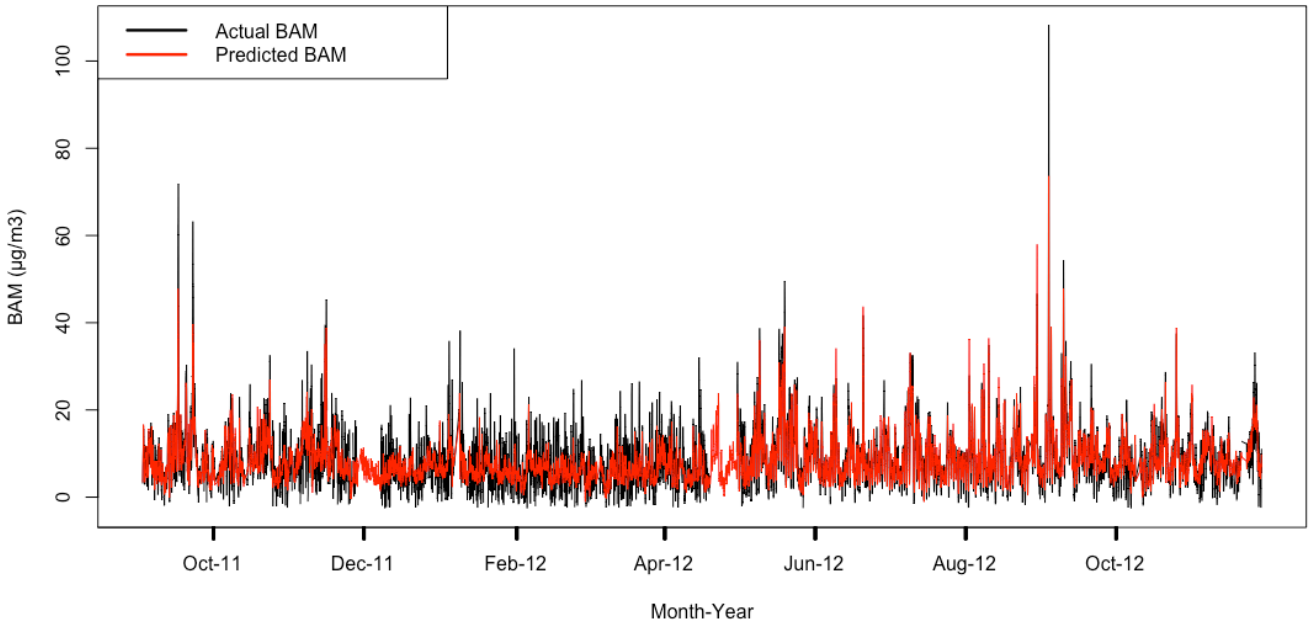
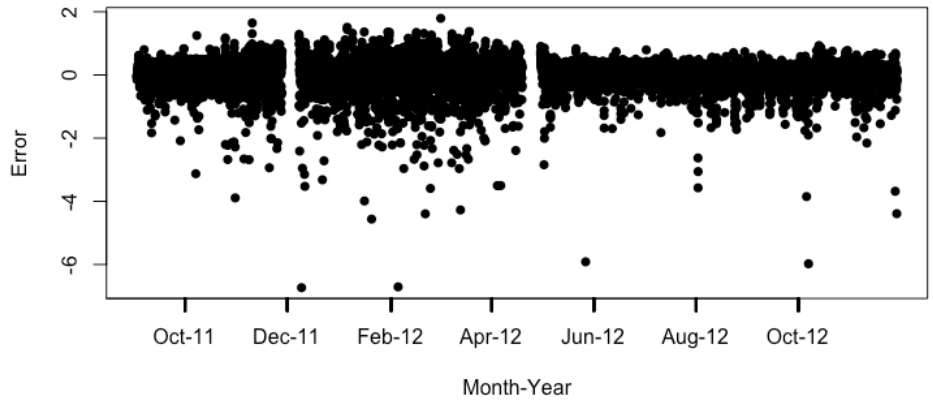


Figure 4- 11. Time series of actual BAM (black) and predicted BAM (red) values over the period of predictions made using the time-series cross validation. The BAM values have been converted back to µg/m^3.

A time series plot of hourly data for the predictions made by the time-series cross validation show that the model does not capture the extreme cases well, with fitted values (red) underestimating the actual BAM readings (black) for most peaks (Figure 4- 11). Especially between November 2011 to May 2012, the modelled values under predict the extreme cases (Figure 4- 11). From May 2012 onwards, it appears that the predicted BAM values have a slightly better predictive ability for extreme cases (Figure 4- 11). No trends were extrapolated regarding the predictions around the mean, as the clustering made it difficult to interpret (Figure 4- 11).

A time series of error is shown in Figure 4- 12 A) for the predictions made using the time series cross validation. The error was calculated as actual BAM values (transformed) minus modelled BAM values (transformed). There is a lot of scatter in the error, ranging from -6.97 to 1.87. From the 8,746 BAM values calculated, 4236 (48.32%) of these under predicted the actual BAM value, and 4,531 (51.68%) of these over predicted the actual BAM value. The mean error for the modelled values is 0.43 $\mu g/m^3$, with a standard deviation of 3.93 $\mu g/m^3$. The histogram and frequency plots (Figure 4- 12 B) suggest a normal Gaussian distribution of the error terms, satisfying the normality assumption of the model.

A)



B)

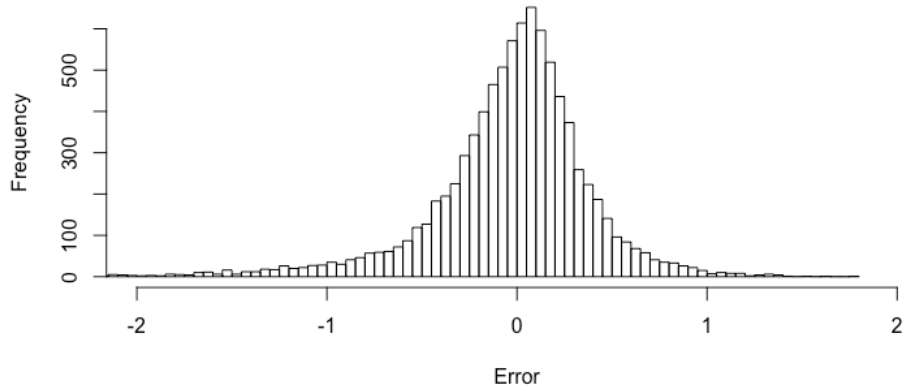


Figure 4- 12. Distribution of error. A) showing a time series and the changes in error, B) showing a histogram of the distribution of error over the period of predictions made using the time series cross validation. The *x*-upper and lower limit is set to +-2, with 68 error values being cut off from the display as they have a value of < -2. The error units are the same as the model, transformed.

The modelled data from the time series cross validation is exhibited in Figure 4- 13. The model under predicts on an hourly basis Figure 4- 13 B). The highest reading of the TEOM (106.9 $\mu g/m^3$) occurred at 2:00 a.m. and the highest reading of the BAM (121.6 $\mu g/m^3$) occurred at 3:00a.m. on the 04/09/2012. The high variance of these predictors within the hours 02:00am to 3:00am may be producing the large confidence interval around this time Figure 4- 13 B). The modelled values track fairly well between 9:00a.m. and 3:00p.m., maintaining a fairly constant difference with the actual BAM values, but then so does the TEOM (Figure 4- 13 B). The modelled hourly data over predicts slightly between 06:00a.m. and 07:00 a.m. (Figure 4- 13 B).

Looking at monthly predictions, the predictive ability of the model is poor between December through to April (Figure 4- 13 C). There is a large discrepancy between the actual and modelled BAM values in these months (~1.0-2.0 $\mu g/m^3$). However, the model does seem to track pretty well between May and November, with only a slight under prediction occurring (Figure 4- 13 C). The model has predicted the months of July very well (Figure 4- 13 C).

On a daily basis, the model under predicts the actual observations on every day by approximately 0.5 - 1.0 $\mu g/m^3$ (Figure 4- 13 D). However, it provides a much closer reading of the actual $PM_{2.5}$ readings than the TEOM does.

Seasonal plots of the hourly (Figure 4- 14 A) and daily (Figure 4- 14 B) shows there are large discrepancies between the actual and predicted BAM values over summer. The model has a poor predictive ability between 6:00a.m. and 9:00p.m for the summer period (Figure 4- 14 A). The daily prediction for summer underestimates actual readings by approximately 1.5 - 2.0 $\mu g/m^3$ (Figure 4- 14 B). On the contrary, the hourly predictions for winter perform very well between 06:00a.m. and midnight (Figure 4- 14 A). On a daily basis, the performance of the modelled overall values is good (Figure 4- 14 B). The hourly and daily data for autumn reveals a fairly constant under prediction in the modelled values (Figure 4- 14 A & B). Observing the hourly data for spring shows a fairly good predictive ability of the model, except between 6:00p.m. and 11:00p.m., where the modelled values have under estimated the true values (Figure 4- 14 A). Daily averages modelled for spring slightly under predict the actual observed values (Figure 4- 14 B).

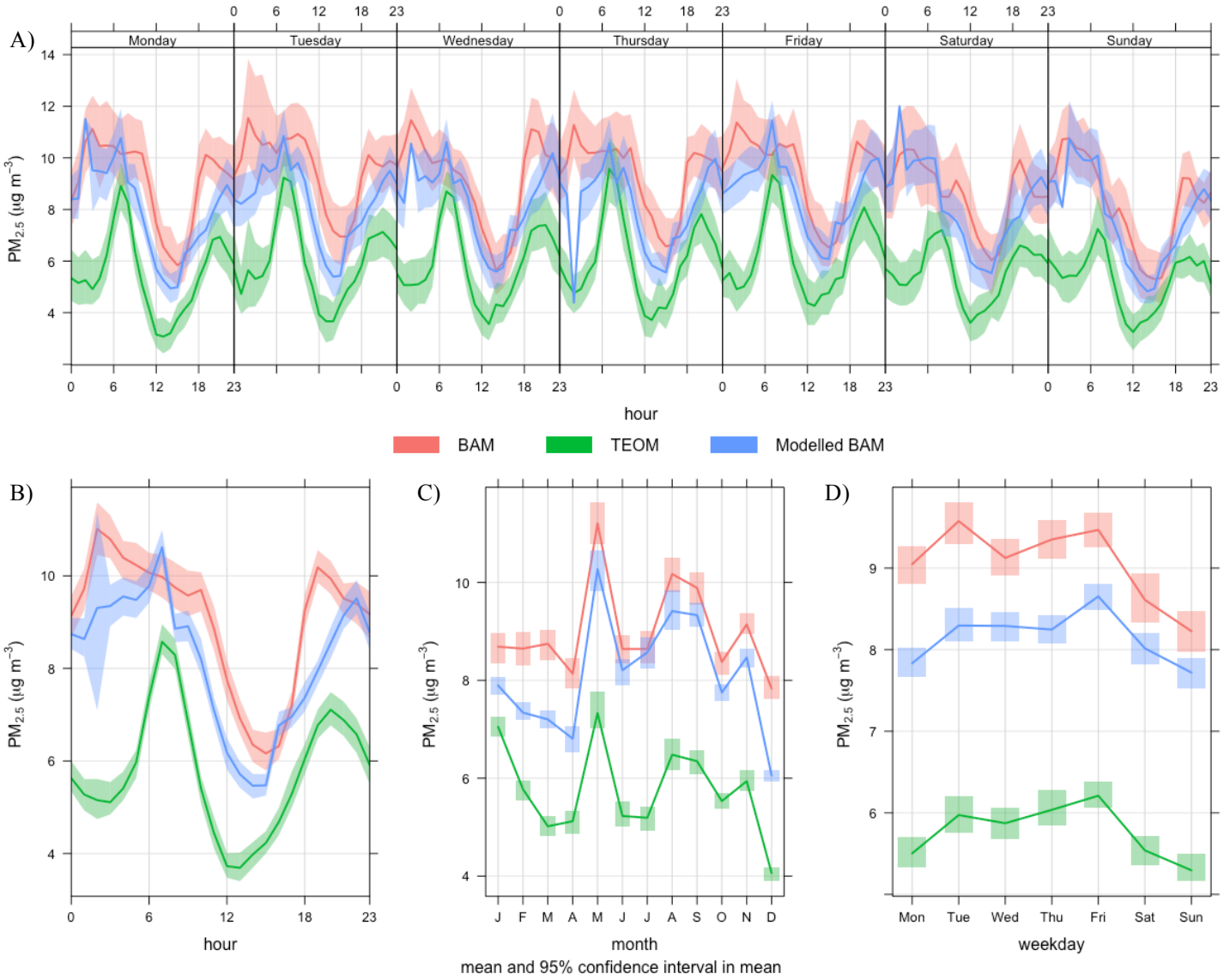


Figure 4- 13. Time Variation plots showing the original BAM (red) and TEOM (green) from the collocated period. The modelled BAM values are indicated by the blue line. A) Hourly-daily, B) Hourly, C) monthly and D) daily plots are shown. The shading around the boxes indicates a 95% confidence interval.

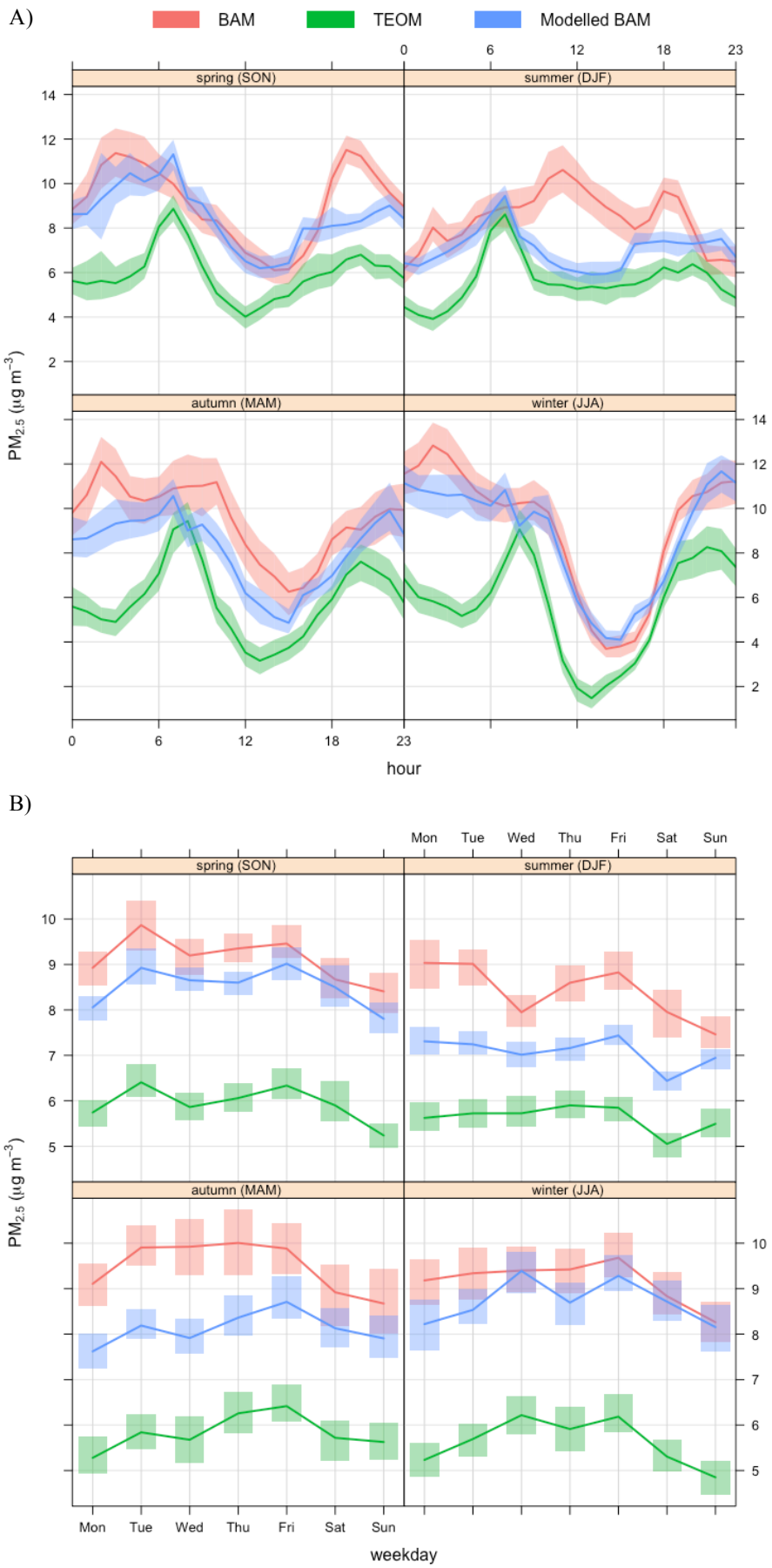


Figure 4- 14. Time variation showing the original BAM (red) and TEOM (green) from the collocated period. The modelled BAM is indicated by the blue line. A) shows hourly data broken up seasonally, and B) shows daily data broken up seasonally. The shading around the boxes indicate a 95% confidence interval.

4.8 Ranking covariates by importance for prediction

It is of great importance to know which variables contribute the most to explaining the predictor variable. Therefore, a statistical measure to rank the variables in terms of their importance was employed. To do this, we recorded the change in the R^2 value when the variable being analysed is added to the model as the last variable. The change in R^2 represents the amount of unique variance that each covariate explains, beyond the other variables in the model. The results are shown in Table 4- 6. The final model, with all variables had an R^2 value of 0.4306, so the difference is calculated as 0.4306 minus the R^2 when the variable is not included in the model. Not surprisingly, NEPH lag 1 was ranked as the most important variable, with a difference in R^2 of 0.0418. One would assume a relationship between the two variables as they are both measuring particles in ambient air, and exhibited a high rho value of 0.59. The NEPH lag 1 is approximately four times as important in predicting BAM than the second and third variables, the time of day and relative humidity. The R^2 improved by 0.0008 when the TEOM lag 24 was removed from the model, although it is strongly significant when included in the overall model, with a *p-value* of 0.00. The NEPH lag 1 is such a crucial variable when predicting BAM, that it alone yields an R^2 of 0.35 with BAM. This is remarkable given that the addition of 13 other variables only improves the R^2 by 7.9%, to 0.43.

4.9 Summary

From the evaluation of the model, through statistics and graphical representations of the data, we can conclude that the hourly models performance is only mediocre at predicting $PM_{2.5}$. The R^2 between the transformed BAM and TEOM is 0.24. Using the ARDL model to correct the $PM_{2.5}$ hourly values considerably improves the R^2 between the BAM and modelled BAM, to 0.43. The predictors ulitized are statistically significant, and do contribute a great amount to calculating the response variable, but it becomes clear through the ACF and PACF plots (see Figure 4- 3), that more variables need to be incorporated into the model if we wish to statistically compute a model with a higher predicative ability, that is more statistically robust.

Although the predictive ability of the model is not strong, the time variation plots indicate that it is a lot more precise simply using the TEOM data (Figure 4- 13 and Figure 4- 14). Therefore, we will move forward into the next chapter, and apply the ARDL model to correct the historical TEOM record for the Chullora site.

Table 4- 6. Table showing each variable, the R^2 value when the particular variable was not included in model, and the difference between the initial model ($R^2 = 0.4306$) and the model with that particular variable excluded. The variables were then ranked in terms of their importance.

Variable	R^2 when variable not included in model	Difference	Rank of importance
NEPH lag 1	0.3888	0.0418	1
Hour block	0.4202	0.0104	2
RH	0.4206	0.0100	3
CO	0.425	0.0056	4
TEMP	0.4252	0.0054	5
TEOM lag 2	0.4258	0.0048	6
Wind Speed	0.4265	0.0041	7
NO	0.4266	0.0040	8
Month Block	0.4279	0.0027	9
TEOM	0.4282	0.0024	10
PM10 lag 2	0.4291	0.0015	11
PM10 lag 1	0.4292	0.0014	12
TEOM lag 1	0.4295	0.0011	13
NO2	0.4298	0.0008	14
TEOM lag 24	0.4314	-0.0008	15

Chapter 5: Application.

5.1 Overview

In this chapter, the ARDL model developed in Chapter 4 is applied to measurements at the Chullora site, from 23/01/2004 at 1:00 a.m. through to 29/11/2012 at 12:00 a.m. For this study the model was applied to values for which all covariates were available, a total of 49,687 observations over the ~9-year period.

5.2 Application of hourly model

As established previously in Chapter 4, the modelled BAM does under predict the actual BAM over the collocated period, with a satisfactory agreement between the actual and predicted BAM concentrations for the collocated period ($R^2 = 0.44$). Looking beyond the collocated period, it is difficult to extrapolate any trend from the time series data, as the hourly data points are so dense (Figure 5- 1). However, examining some statistics can assist with this.

The summary statistics presented in Table 5- 1 serve as a useful tool to get an idea of how the distribution of $PM_{2.5}$ in the period 2004 to 2012 may have changed using the predictions from our ARDL model. The year with the highest hourly mean of $PM_{2.5}$ is 2004, at 10.08 $\mu g/m^3$, with a median of 8.78 $\mu g/m^3$, and a standard deviation of 6.09 $\mu g/m^3$. However, this year has a data capture of only 66.02%, with missing values likely influencing these results. The mean hourly prediction drops to 9.45 $\mu g/m^3$ in 2005, then slightly rises again in 2006, to 9.49 $\mu g/m^3$. From 2007, to 2009, the predicted BAM fluctuate, leading to the lowest mean year in 2010, with a prediction of 7.95 $\mu g/m^3$. The modelled $PM_{2.5}$ rise in 2011 and 2012 to 8.39 $\mu g/m^3$ and 8.21 $\mu g/m^3$ respectively. The minimum predicted value and standard error remains fairly constant over the years of the study period. Compared with the actual values displayed at the bottom part of the table, the modelled values of 7.95 $\mu g/m^3$, 8.39 $\mu g/m^3$, and 8.21 $\mu g/m^3$ underestimate the actual BAM readings of 8.50 $\mu g/m^3$, 9.57 $\mu g/m^3$ and 8.72 $\mu g/m^3$ for 2010, 2011 and 2012 respectively.

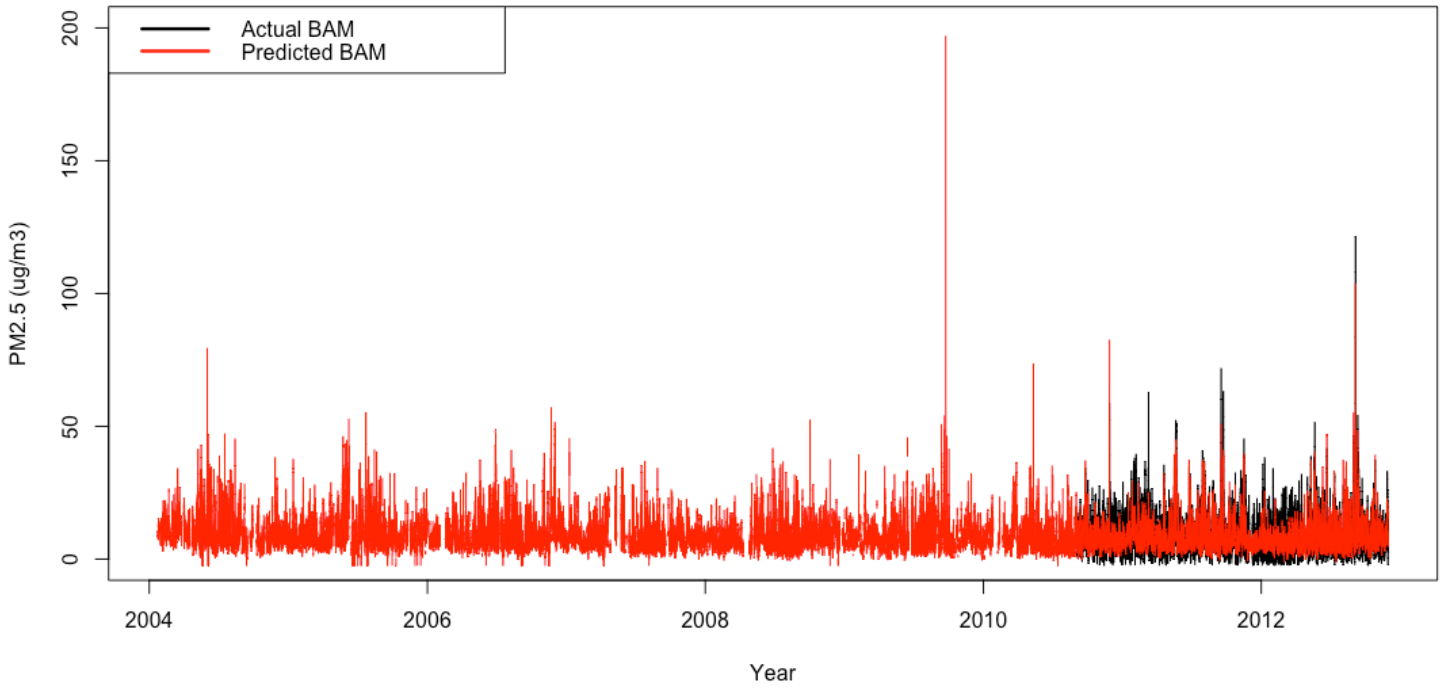


Figure 5- 1. Hourly time series of actual BAM (black) and modelled BAM (red) for the period from 23/01/2004 to 29/11/2012.

Table 5- 1. Summary statistics for BAM predictions made from 2004 to 2012.

Year	Number of observaions	Mean ($\mu g/m^3$)	Standard deviation ($\mu g/m^3$)	Median ($\mu g/m^3$)	Min ($\mu g/m^3$)	Max ($\mu g/m^3$)	Standard error ($\mu g/m^3$)	NA's	% of data missing
Modelled values									
2004	5450	10.08	6.09	8.78	-2.51	79.48	0.08	2805	33.98
2005	5887	9.45	5.98	8.06	-2.51	55.14	0.08	2873	32.80
2006	6286	9.49	5.73	8.27	-2.51	57.15	0.07	2474	28.24
2007	6285	8.10	4.97	7.07	-2.51	45.52	0.06	2475	28.25
2008	5795	8.47	5.07	7.51	-2.51	52.40	0.07	2989	24.78
2009	5987	8.67	7.19	7.27	-0.80	207.10	0.09	2773	31.66
2010	6395	7.95	5.11	6.88	-2.51	82.61	0.06	2365	27.00
2011	7485	8.39	5.12	7.24	-0.61	50.93	0.06	1275	14.55
2012	6874	8.21	5.26	7.07	-1.22	103.90	0.06	1119	14.00
Actual values									
2010	2816	8.50	5.18	7.90	-2.5	58.6	0.10	71	2.46
2011	8240	9.57	6.87	8.30	-2.5	71.8	0.08	520	5.94
2012	7592	8.72	6.78	7.60	-2.5	121.6	0.08	401	5.01

Note: Actual values were recorded from 02/09/2010 to 29/11/2012 so 2010 and 2012 actual values do not represent a whole year's worth of data.

Time variation plots of the actual and predicted BAM values broken down by year from 2004 to 2012 are shown in Figure 5- 2. In general, the $PM_{2.5}$ hourly readings over the day (Figure 5- 2 A) from 2004 to 2012 follow a similar pattern. There is a peak around 7:00 a.m., with a drop throughout the day, then rising again around 4:00 p.m. coinciding with afternoon traffic. This rises until midnight, where some years' experience a fall and some plateau in $PM_{2.5}$ from midnight till 6:00 a.m. Based on our predictive model, it seems that average hourly $PM_{2.5}$ levels were highest in 2004, and slowly decrease over the years (Figure 5- 2 A). The years of 2004, 2008 and 2012 exhibit a strong trough during the day at 2:00p.m. at values of approximately 5.00 – 7.00 $\mu g/m^3$(Figure 5- 2 A). Examining the summer of 2010/11 and 2011/12, observed BAM values are drastically different (Figure 5- 2 B). In summer 2010/11, BAM readings are higher, at 9.00-10.50 $\mu g/m^3$, compared to summer 2011/12, where actual BAM readings are lower, at approximately 7.00 $\mu g/m^3$ (Figure 5- 2 B). This highlights the variation in monthly trends, emphasising how complex the processes are that dominate the summer $PM_{2.5}$ readings, i.e. semi-volatiles. The predictive model does pick up these differences fairly well. Based on our collocated results, we assume that average monthly modelled $PM_{2.5}$ values from 2004 to 2012 (Figure 5- 2 B) consistently under predict the actual BAM.

`TheilSen` is a function from the *Openiar* package in R, which aids in determining percentage changes in $PM_{2.5}$ in our predictions from 2004 to 2012. According to our predictions between 2004 and 2012, the decrease in $PM_{2.5}$ between 2004 and 2012 is statistically significant, with a decrease of 2.25 % per year, with a 95% confidence interval of -2.97%, -1.27% Using the modelled (2004 to 2010) and actual (2010 to 2012) values, there is a statistically significant 1.72% decrease per year in $PM_{2.5}$ concentrations (95% confidence interval of -2.6% and -0.71%) (Figure 5- 3). This further supports the idea that the hourly model underestimates actual $PM_{2.5}$ concentrations. Seasonally, the change in distribution of modelled (2004 to 2010) and actual (2010 to 2012) $PM_{2.5}$ is statistically significant in spring, summer and winter, with an increase in spring of 2.43% per year (1.48%, 3.45%), a decrease in summer of -2.52% per year (-3.38%, 1.67%), and a decrease in winter of -2.98% per year (-3.77%,-2.38%) (Figure 5- 4).

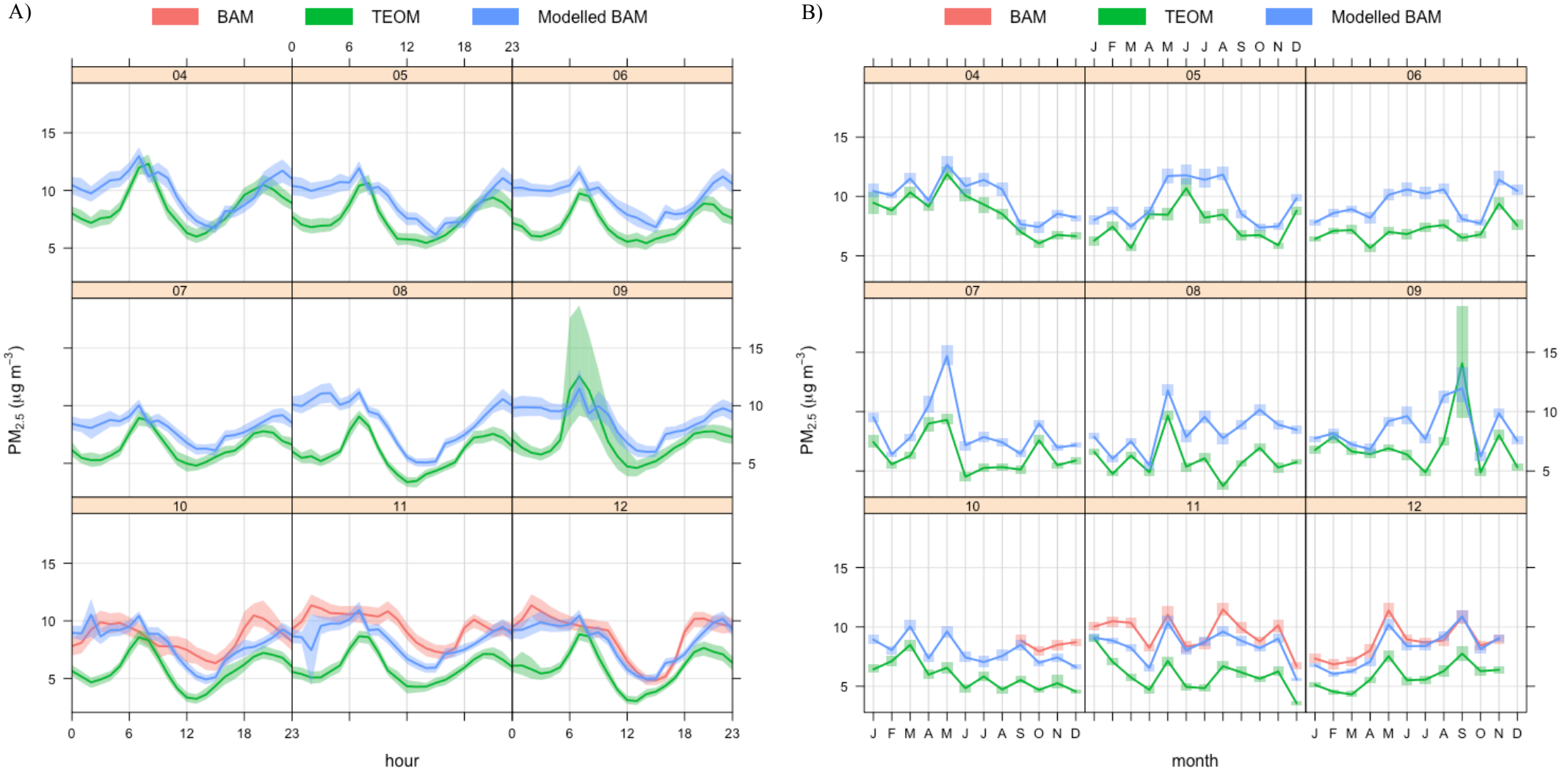


Figure 5- 2. Time variation plots of the actual BAM readings (red), the actual TEOM readings (green) and the modelled BAM readings (blue) from 2004 to 2012, with A) showing the variation at an hourly time scale and B) showing the variation at a monthly time scale.

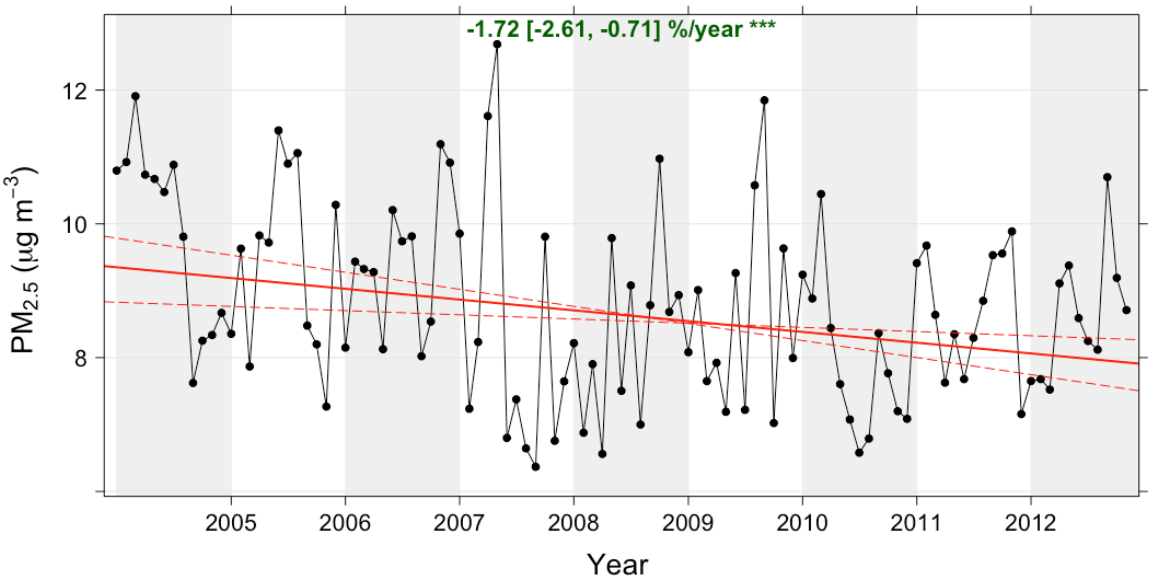


Figure 5- 3. Change in $PM_{2.5}$ from 2004 to 2012 based on the modelled (2004 to 2010) and actual (2010 to 2012) values. Also shown is the average % decrease in $PM_{2.5}$ per year with 95% confidence intervals. The three green stars indicates the change in $PM_{2.5}$ over the year is statistically significant.

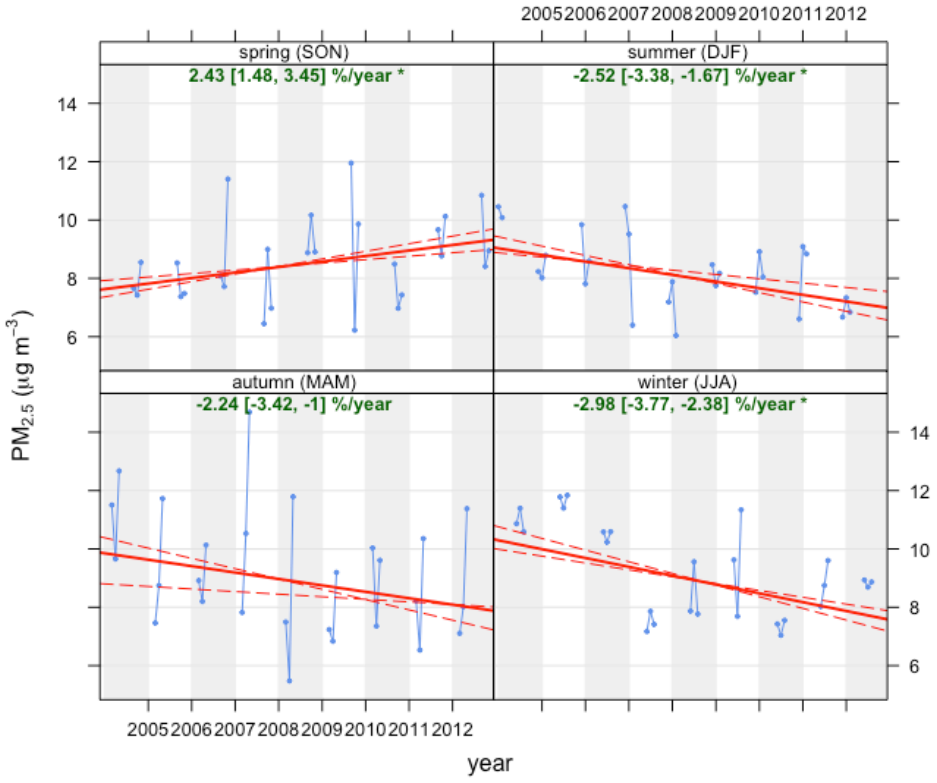


Figure 5- 4. Change in $PM_{2.5}$ from 2004 to 2012 shown seasonally, based on the predicted (2004 to 2010) and actual (2010 to 2012) values. Also shown is the average % decrease or increase in $PM_{2.5}$ per year with 95% confidence intervals. The green stars indicate the change in $PM_{2.5}$ over the seasons per year is statistically significant (spring, summer and winter).

5.3 Summary

The results suggest a gradual decrease in average hourly $PM_{2.5}$ levels from 2004 to 2012, of 1.72% per year (-2.61%, -0.71%). However, the high percentage of missing data from 2004 to 2010, ranging from 24.78-33.98% may influence trend calculation. Seasonally, there is a statistically significant decrease in $PM_{2.5}$ in spring, summer and winter between 2004 to 2012, of 2.43% (1.48%, 3.45%) increase per year in spring, -2.52% (-3.38%, -1.67%) per year in summer and -2.98%(-3.77%, -2.38%) per year in winter.

Of more relevance to this study is the models performance. Given that the hourly ARDL model under-predicts actual values, as demonstrated by the `TheilSen` output and Table 4- 5, we infer that these back-predicted modelled values too underestimate the true BAM values from 2004 to 2012.

Due to the limited predictive of the model, and in attempts to provide the OEH with a better performing one, an ARDL model was developed to predict $PM_{2.5}$ on a *daily* basis. This will assist in reducing any positive and negative biases, while also smoothing out short-term variations and expressing longer-term trends (Li et al., 2012). This is developed in Chapter 6.

Chapter 6: Daily predictive model.

6.1 Overview

Given the limited predictive ability of the hourly predictive model, a daily predictive model was constructed in attempts to provide an improved predictive model, one with more stability by reducing the uncertainty as a result of averaging over many observations.

6.2 Exploratory data analysis

Before building a daily prediction model, we need to ensure that the daily data still meets the assumptions of an ARDL model. To check this, an exploratory data analysis was performed and the assumptions of the model were checked. To prevent this section from being a repeat of the Exploratory Data Analyses presented in Chapter 3, this section instead focusses on recording the changes between patterns of the hourly and daily data.

6.2.1 Available data

As previously outlined, the hourly concentrations were averaged over the 24-hour (1:00 a.m. to midnight) period. Under national air quality guidelines and protocols, days with less than 75% data capture were excluded from the 24-hour averages (Office of Environment & Heritage, 2012). The available data for each of the variables, between 03/09/2010 to 29/11/2012 (819 days), is shown in Table 3- 2 in Chapter 3.

6.2.2 Comparisons of measurements from the collocated TEOM and BAM

The raw daily TEOM and BAM readings for the collocated period agree well (Figure 3- 1; raw $R^2 = 0.80$). There is still some scatter along the least squares regression line, but a lot less than the hourly data (Figure 3- 1). Most of the raw daily TEOM data is bound by 2 $\mu g/m^3$ and 10 $\mu g/m^3$ and the BAM is bound by 3 $\mu g/m^3$ and 15 $\mu g/m^3$ (Figure 3- 1).

Table 6- 1. Instrument inter-comparison through basic statistics (based on daily data).

	TEOM ($\mu g/m^3$)	BAM ($\mu g/m^3$)
Mean	5.77	9.04
Median	8.25	5.05
Standard Deviation	3.28	4.08
Inter quartile range	3.70	4.60

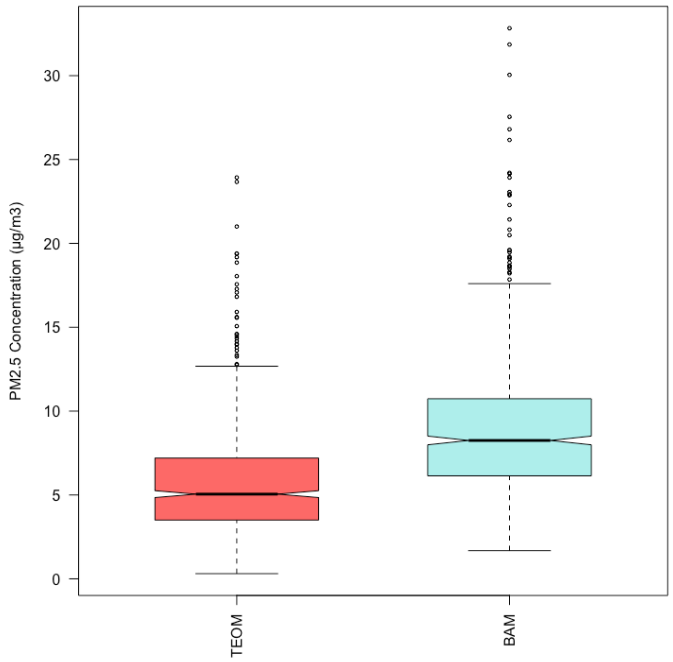


Figure 6- 1. Box and whisker plot showing TEOM and BAM measurements, based on daily averages, for the collocated period.

There is a definite difference between the daily BAM and TEOM measurements (Figure 6- 1 and Table 6- 1); the median of the BAM is higher than the TEOM (8.25 µg/m^3 and 5.05 µg/m^3 respectively), with the BAM having a larger interquartile range (IQR BAM = 4.60 µg/m^3, IQR TEOM = 3.70 µg/m^3) . There are some outliers in the daily data, as indicated by the dots that extend beyond the whiskers in Figure 6- 1.

The TEOM and BAM readings still display a large discrepancy in their values, at a daily and monthly scale (Figure 6- 2). The difference in daily $PM_{2.5}$ remains constant, and is not dependent on the day of the week (Figure 6- 2 A). However, there are changes in the difference of $PM_{2.5}$ depending on the month, with the warmer months of January and February reading closer than the remaining months, with these months maintaining a fairly constant difference in their readings of $PM_{2.5}$ (Figure 6- 2 B).

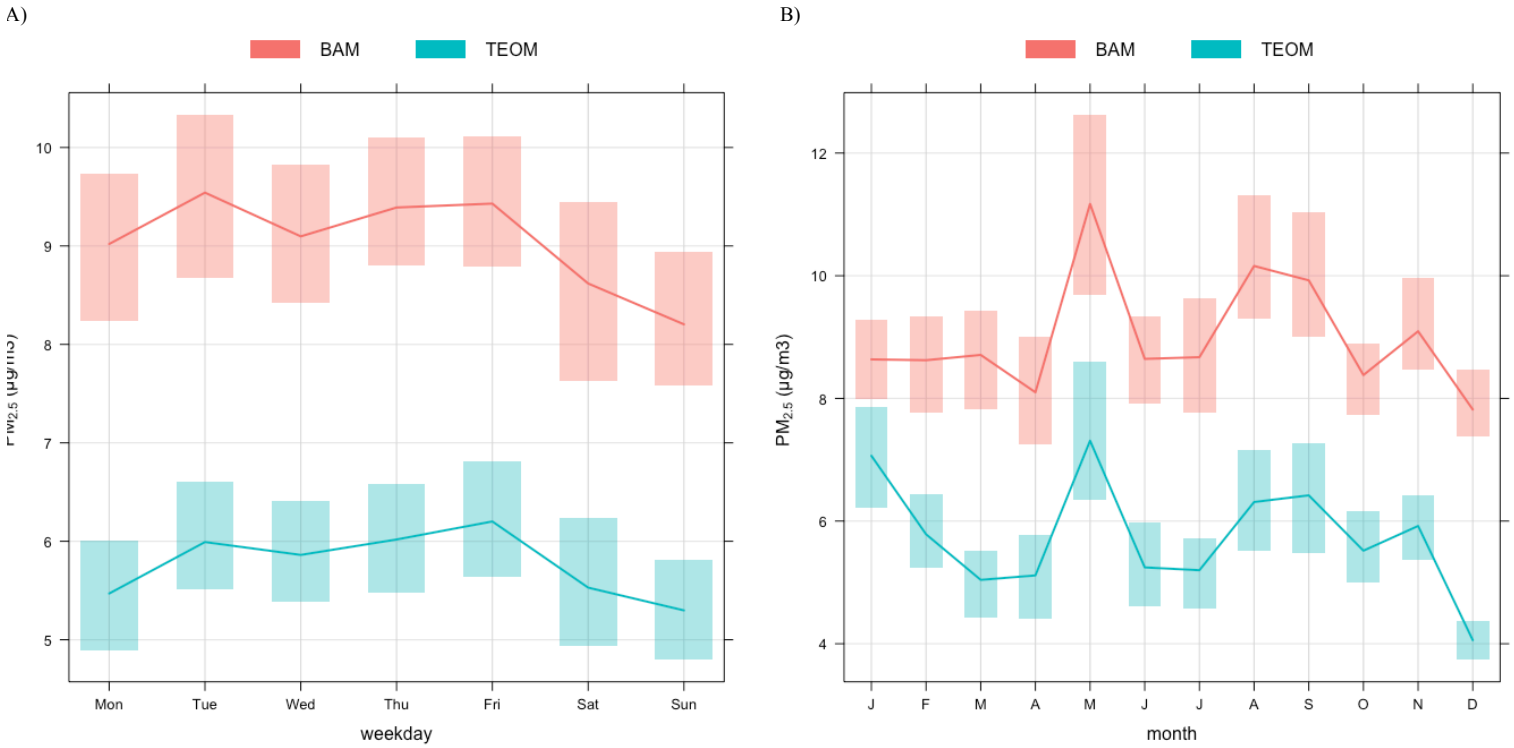


Figure 6- 2. Time variation plot for daily data from the collocated period, for BAM (red) and TEOM (blue). The lines show a 95% confidence interval. A) shows the daily data broken up per day of the week, and B) shows the daily data averaged per month.

6.2.3 Correlation of $PM_{2.5}$ BAM with other variables

The descriptive statistics of all air quality parameters and meteorological conditions are shown in Table 6- 2. There were no negative minimum values for BAM, TEOM, PM_{10} and Nephelometer data. Maximum daily averages occurred in spring for both BAM (32.8 µg/m^3), TEOM (23.9 µg/m^3) and Nephelometer data (1.6 bsp) (Table 6- 2), on the same day (04/09/12), suggesting a strong link between these variables. Corresponding results were observed for temperature, relative humidity, carbon monoxide, nitrogen monoxide, nitrogen oxides, nitrogen dioxide, sulfur dioxide, ozone, wind speed, wind direction and variation in wind direction (Table 6- 2).

Table 6- 2. Descriptive statistics for air pollution and meteorological parameters, shown seasonally, based on daily data.

Parameter	Season	Variance	Minimum value	Maximum value	Standard deviation	Mean
$PM_{2.5}$ (BAM) ($\mu g/m^3$)	Autumn	24.6	2.7	31.9	5.0	9.5
	Spring	17.3	2.6	32.8	4.2	9.1
	Summer	8.2	2.6	20.5	2.9	8.4
	Winter	16.2	1.7	24.2	4.0	9.2
$PM_{2.5}$ (TEOM) ($\mu g/m^3$)	Autumn	13.9	0.3	23.7	3.7	5.8
	Spring	11.9	0.6	23.9	3.4	5.9
	Summer	8.4	1.1	19.2	2.9	5.6
	Winter	8.2	1.2	15.6	2.9	5.6
PM_{10} (TEOM) ($\mu g/m^3$)	Autumn	69.5	5.7	64.5	8.3	18.2
	Spring	56.8	4.3	53.5	7.5	19.1
	Summer	56.5	4.9	58.5	7.5	18.1
	Winter	73.7	6.4	58.8	8.6	18.7
Nephelometer (bsp)	Autumn	0.0	0.0	1.4	0.2	0.3
	Spring	0.0	0.0	1.6	0.2	0.3
	Summer	0.0	0.1	0.7	0.1	0.2
	Winter	0.0	0.1	1.0	0.2	0.3
Temperature (°C)	Autumn	12.2	10.2	25.5	3.5	17.5
	Spring	11.8	10.1	26.7	3.4	17.5
	Summer	8.8	14.5	33.2	3.0	22.1
	Winter	3.2	8.8	19.2	1.8	12.3
Relative Humidity (%)	Autumn	116.7	46.7	97.9	10.8	73.7
	Spring	134.9	27.2	95.1	11.6	66.7
	Summer	94.9	35.7	96.2	9.7	71.9
	Winter	168.7	41.5	97.6	13.0	70.6
Carbon Monoxide (ppm)	Autumn	0.0	0.2	0.9	0.1	0.4
	Spring	0.0	0.1	0.7	0.1	0.3
	Summer	0.0	0.0	0.5	0.1	0.3
	Winter	0.0	0.1	0.9	0.2	0.4
Nitrogen monoxide (ppb)	Autumn	331.6	-0.4	79.2	18.2	18.3
	Spring	76.8	-0.6	51.7	8.8	9.2
	Summer	30.1	-0.4	38.6	5.5	5.9
	Winter	410.4	-0.3	106.8	20.3	22.9
Nitrogen Oxides (ppb)	Autumn	488.2	5.8	101.9	22.1	32.2
	Spring	161.2	3.6	78.5	12.7	22.7
	Summer	56.0	3.3	47.5	7.5	14.9
	Winter	580.6	5.5	129.0	24.1	39.3
Nitrogen Dioxide (ppb)	Autumn	21.9	5.6	26.7	4.7	13.8
	Spring	23.2	3.6	27.0	4.8	13.4
	Summer	9.0	2.8	17.8	3.0	8.8
	Winter	21.7	5.3	28.9	4.7	16.3
Sulfur Dioxide (ppb)	Autumn	0.0	-0.1	0.4	0.1	0.1
	Spring	0.0	-0.1	0.5	0.1	0.1
	Summer	0.0	0.0	0.3	0.1	0.1
	Winter	0.0	-0.1	0.4	0.1	0.1
Ozone (ppb)	Autumn	0.2	0.1	2.1	0.4	1.1
	Spring	0.2	0.3	3.6	0.5	1.7
	Summer	0.3	0.2	3.7	0.5	1.4
	Winter	0.3	0.1	2.7	0.5	1.0
Wind Speed (m/s)	Autumn	0.5	0.7	4.9	0.7	1.9
	Spring	0.5	0.8	5.7	0.7	2.1
	Summer	0.5	0.9	5.0	0.7	2.2
	Winter	0.8	0.7	4.9	0.9	2.0
Wind Direction (°)	Autumn	2318.0	56.1	299.4	48.1	202.5
	Spring	2733.5	45.7	296.8	52.3	173.5
	Summer	2361.2	40.7	293.8	48.6	152.1
	Winter	1413.2	148.5	313.5	37.6	230.9

6.2.4 Transforming data

Transforming for symmetry

Given the BAM, $PM_{2.5}$ TEOM, PM_{10} TEOM and NEPH did not have any negative values, the transformation method used is a straight logarithm. The symmetry of the data sets were then examined to determine if the symmetry improved, which they did for all four variables (AP9-1in Appendix 9). The linear regression of the transformed TEOM and BAM is shown in Figure 6- 3, with the R^2 equalling 0.75.

The same method used adjust the hourly gases was applied to the daily data on the remaining gasses (NO_x, NO, NO_2, SO_2, Ozone), that is:

$$x = \log(y_t + 1),$$

where x is the transformed value of the gas and y is the raw gas concentration at time t. The symmetry of NO_x, NO_2, SO_2 and Ozone all improved when transformed (Appendix 9).

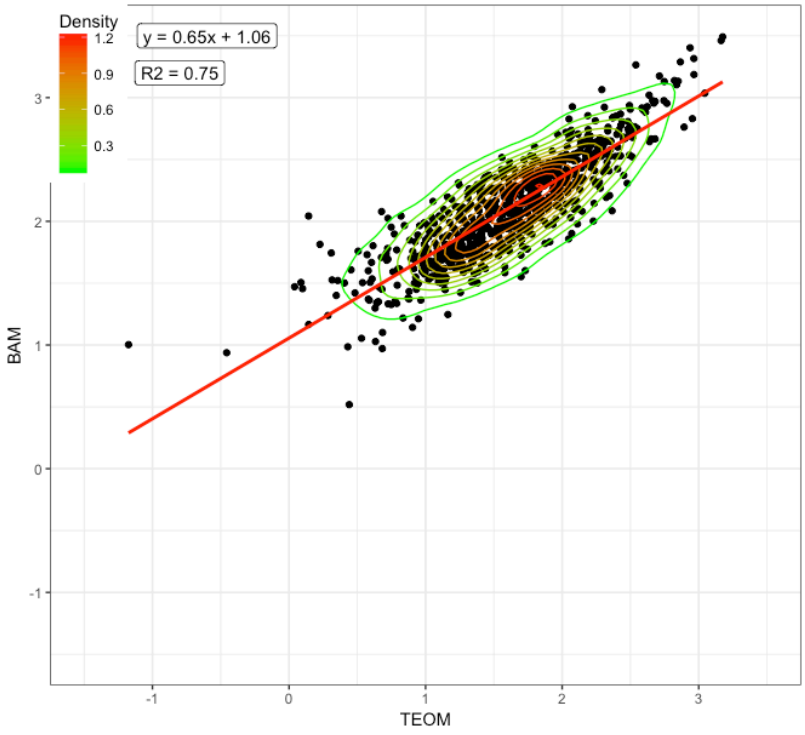


Figure 6- 3. Scatterplot showing density of daily averaged points for the transformed TEOM and BAM over the collocated period. The least squares regression line (red), equation for the line, and R^2 value is displayed.

Transforming for straightness

The linearity of all variables, except for Ozone and SO_2, improved once transformed (Figure AP9-2, Figure AP9-3 and AP9-4 in Appendix 9).

6.2.5 Lagged variables

Lagged independent and dependent variables influence the dependent variable (Table 6- 3). Variables with a lag of 1 may be useful in the prediction model, but lag 2 variables were not as useful, all possessing a rho of < 0.25 (Table 6- 3). Therefore, the independent variables with a lag response of 2 were omitted from the pool of variables available for selection for the model.

Table 6- 3. Correlations between $PM_{2.5}$ BAM (time 0) and the independent variables, including their lagged values, based on daily values. Lags are at daily intervals.

Instrument	Independent variable	Correlation (Rho value)
$PM_{2.5}$ BAM	Lag 1	0.51
	Lag 2	0.21
$PM_{2.5}$ TEOM	Lag 0	0.86
	Lag 1	0.59
	Lag 2	0.24
PM_{10} TEOM	Lag 0	0.72
	Lag 1	0.45
	Lag 2	0.19
NEPHELOMETER	Lag 0	0.84
	Lag 1	0.55
	Lag 2	0.21

6.2.6 Stationarity

A Ljung-Box test was applied to test the stationarity of the time series using the transformed independent values. The results from the Ljung-Box test suggest that the time series is non-stationary, as the *p-value* is 0.00. The ACF and PACF plots depict a reasonably stationary time series for the TEOM and BAM terms (Figure 6- 4). The ACF and PACF drop within the 95% limit within a few lags, and are a lot less auto correlated than the hourly data (see Figure 3- 11). There is slightly more autocorrelation in the BAM time series than the TEOM time series.

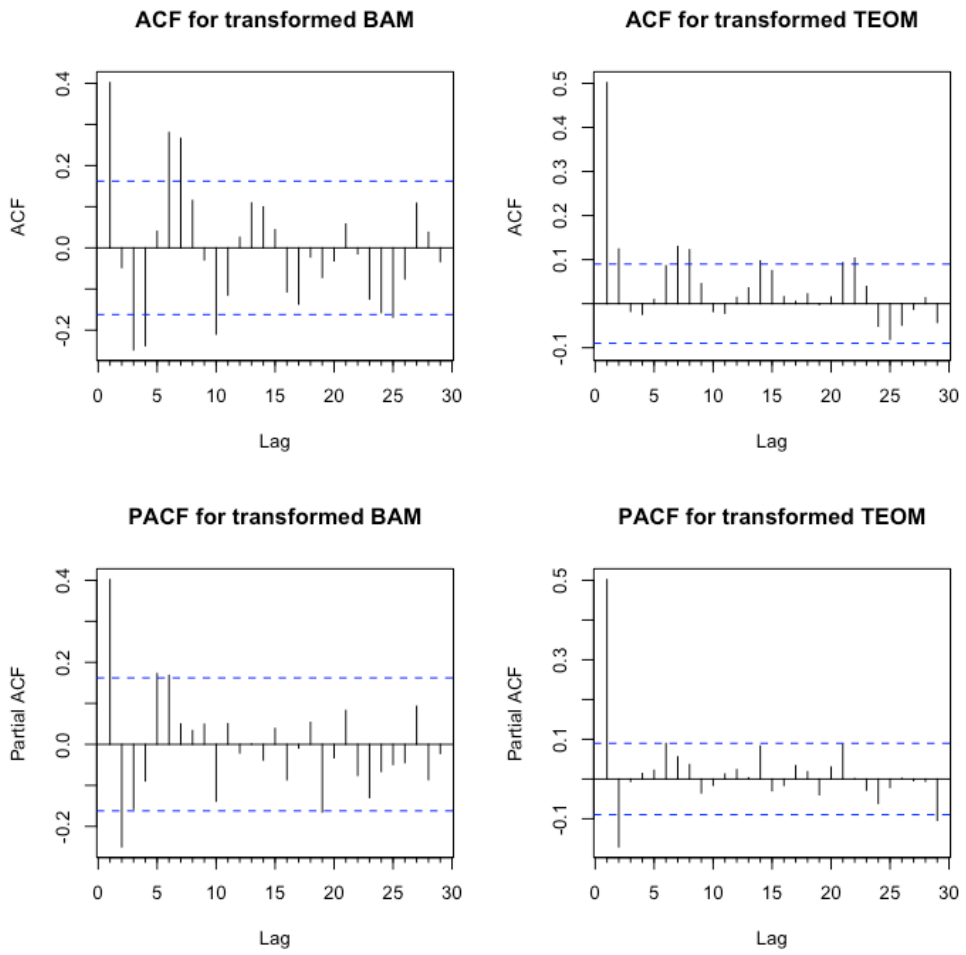


Figure 6- 4. ACF and PACF for transformed daily TEOM and BAM for the collocated period. The lags are at a daily time scale.

6.2.7 Decisions and assumptions of model

As for the hourly predictive model, as ARDL model seemed appropriate for this particular data set. However, we must check that the assumptions of the model are met.

Firstly, the relationship between the independent and dependent variables must be linear. This was previously discussed in *6.2.4 Transforming data.* Variables that do not meet this assumption include Ozone and SO_2.

Next, the model assumes that there is little or no multicollinearity present between variables in the data. A correlation matrix for all variables is shown in Table AP9-1 in Appendix 9. Results demonstrate a high correlation between a number of variables; TEOM and PM_{10} (rho = 0.78), TEOM lag 1 and PM10 lag 1 (rho = 0.78), NEPH and TEOM (rho = 0.86), TEOM lag 1 and NEPH lag 1(rho = 0.86), CO and NO_x (rho = 0.79), CO and NO (rho = 0.77), NO_x and NO (rho = 0.95), and NO_x and NO_2 (rho = 0.94). A decision needs to made on which variables to include and which to disregard. If we wish to keep TEOM and TEOM

lag 1, we would have to exclude PM_{10}, PM_{10} lag 1, NEPH and NEPH lag 1. This did not seem like a wise choice, as we lose 4 predictor variables. Instead, TEOM and TEOM lag 1 were omitted from the pool of variables, simply because this maximised the number of available predictors. Additionally, NO_x and NO were omitted, meaning that CO and NO_2 were available for use.

Next, there must be little or no autocorrelation in the residuals of the model. Also, an ARDL model assumes homoscedasticity. And lastly, the model assumes that the residuals are normally distributed. These three assumptions will be tested in the next section, once the model has been constructed.

6.3 Model building and evaluation

6.3.1 Data preparation

The model was constructed from cases where all covariates were recorded. No outliers were detected through a visual inspection or using a plot showing Cook's distance.

6.3.2 Variable selection and model construction

The variables that are available for selection are shown in Table 6- 4. Months were grouped into significant blocks, for the sake of producing a parsimonious model. Block *a* included months 10 through to 4, and block *b* consisted of months 5 through to 9. Reasons for cut-offs for these blocks is explained in Appendix 10. Note that Ozone and SO_2 were omitted for their non-linearity, and TEOM, TEOM lag 1, NO_x and NO were omitted due to their collinearity with other variables.

Same as for the hourly data, two methods for variable selection were applied, manual forward and backwards f-test selection, and four measures of predictive ability were used to determine the best predictive model; $\bar{R}^2$, CV, AIC and BIC.

Ultimately, both the manual forward and backward f-test selection produced the same model. This emphasises the robustness of the variable selection process. The measures of predictive ability are shown in Table 6- 5. The $\bar{R}^2$ value is strong, at 0.80. The final model summary is shown in Figure 6- 5. All variables are strongly significant. The model is built on 798 complete observations.

Table 6- 4. Table of variables available to be used in the daily predictive model.

Variable name	What is it?
temp	Temperature (°C)
rh	Relative Humidity (%)
mthbk	Monthly data broken into blocks based on significance levels
nephl	Log(neph)
nephl.l1	Lag1(nephl)
pm10l	Log(pm10)
pm10l.l1	Lag1(pm10l)
lco	Log(COppm + 1)
lno2	Log(NO2ppm + 1)
ws	Wind speed (m/s)
wdir	Wind direction (categorical based on °)

Table 6- 5. Results from two methods of variable selection. Measures of predictive ability are shown by the CV, AIC, BIC and $\bar{R}^2$.

CV	AIC	BIC	$\bar{R}^2$
0.03	-2574.23	-2532.49	0.80

```
Call:
lm(formula = baml ~ mthbk + rh + nephl + nephl.l1 + pm10l + lco +
    ws, data = daily)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83529 -0.10727  0.00501  0.11664  0.79346

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.0412948  0.1552257  13.150  < 2e-16 ***
mthbkb      -0.0503122  0.0148134  -3.396 0.000719 ***
rh          -0.0049165  0.0008416  -5.842 7.67e-09 ***
nephl        0.3947454  0.0277023  14.250  < 2e-16 ***
nephl.l1     0.0928386  0.0162299   5.720 1.53e-08 ***
pm10l        0.2328416  0.0296884   7.843 1.50e-14 ***
lco          1.5462651  0.1337178  11.564  < 2e-16 ***
ws           0.0443748  0.0118248   3.753 0.000188 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1839 on 755 degrees of freedom
  (56 observations deleted due to missingness)
Multiple R-squared:  0.8059, Adjusted R-squared:  0.8041
F-statistic: 447.8 on 7 and 755 DF,  p-value: < 2.2e-16
```

Figure 6- 5. Model summary/output for predicting BAM daily values.

6.3.3 Examining residuals

Examination of residuals can tell us a lot about the model and the data, with a good model containing few patterns in the residuals. The results suggest that there is only a small amount of autocorrelation in the residuals, as the lags pass the blue dotted 95% confidence line a few times by a small amount (Figure 6- 6). These exceedances are not enough to be concerned about. The residuals are a lot less auto-correlated than those of the hourly data (see Figure 4- 3).

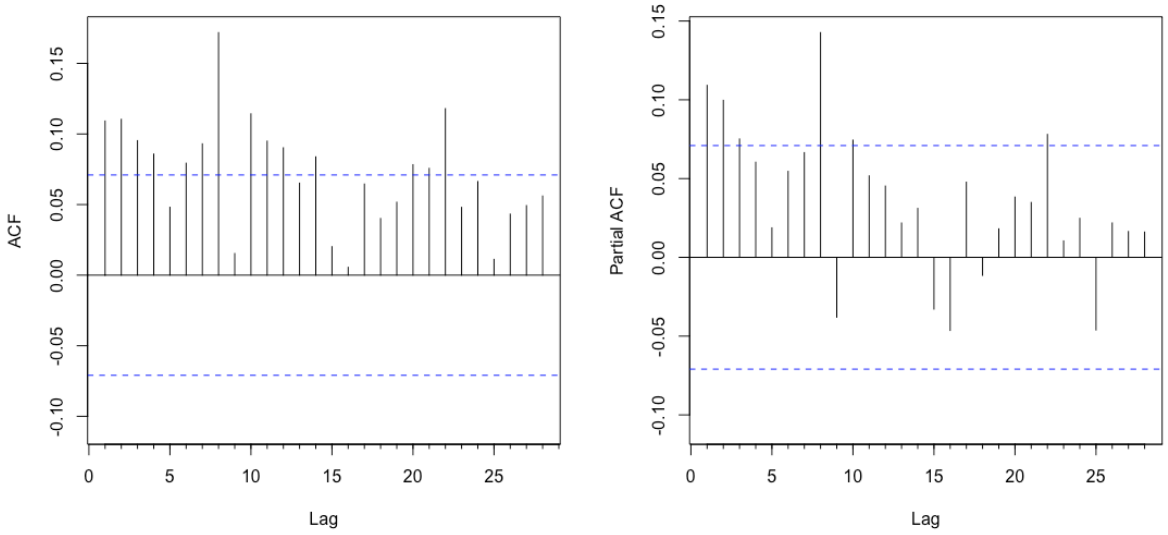


Figure 6- 6. ACF and PACF plots of residuals of final model used for prediction of daily BAM values.

The CCF plots indicate that most of the time, non-stationarity is not induced in the model from the time series structure of the predictors (Figure 6- 7). Most of the time, the lags sit within the 95% limit. Compared to the hourly CCF plots (Figure 4- 5 and Figure 4- 6), the daily plots show more stationarity for the independent variables (Figure 6- 7).

6.3.4 Testing remaining assumptions of model

Testing for homoscedasticity

The residuals occur randomly around the zero line (Figure 6- 8), indicating the suitability of assuming a linear relationship. The residuals roughly form a horizontal band around the zero line (Figure 6- 8), suggesting the variances of the error terms are equal. Lastly, no one residual stands out from the pattern of residuals (Figure 6- 8), suggesting there are no outliers in the data set. All of these indicate homoscedasticity of the daily model.

Testing for normality of the residuals

Figure 6- 9 A) shows a Q-Q plot of the studentized residuals from a linear model against the theoretical quantiles of a comparison distribution. The residuals of the final daily model are fairly normal. Only the extreme values at either tail lie outside of the 95% confidence interval (Figure 6- 9 A). However, as demonstrated by Lumley et al. (2002), such deviations are not a concern for large datasets, and will not affect the outcome of the model. The histogram of the residuals suggests normality Figure 6- 9 B). Therefore, we conclude that the models residuals are normally distributed.

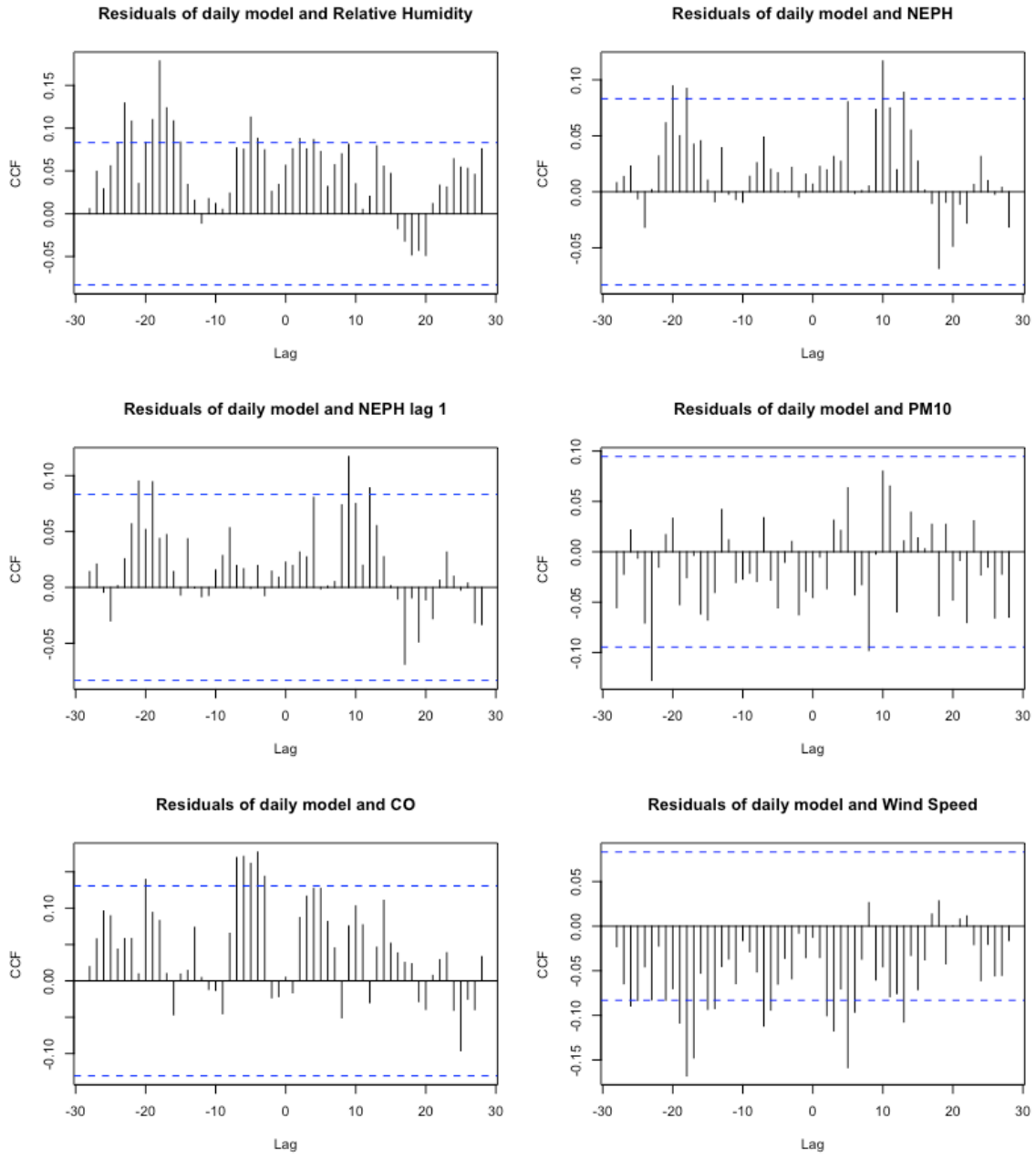


Figure 6- 7. CCF plots for covariates included in the daily model.

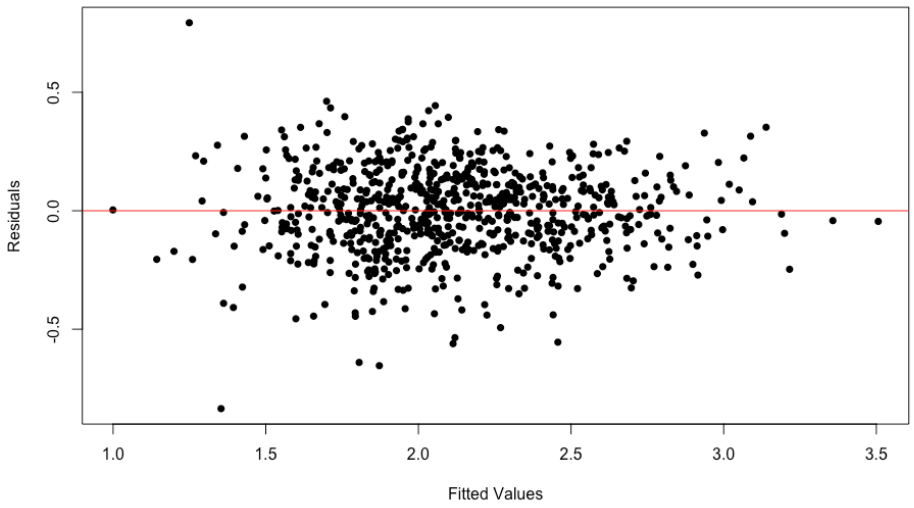


Figure 6- 8. Plot of residuals vs fitted values for the final daily model. The red line has a slope of 0 along the *y*-intercept of 0.

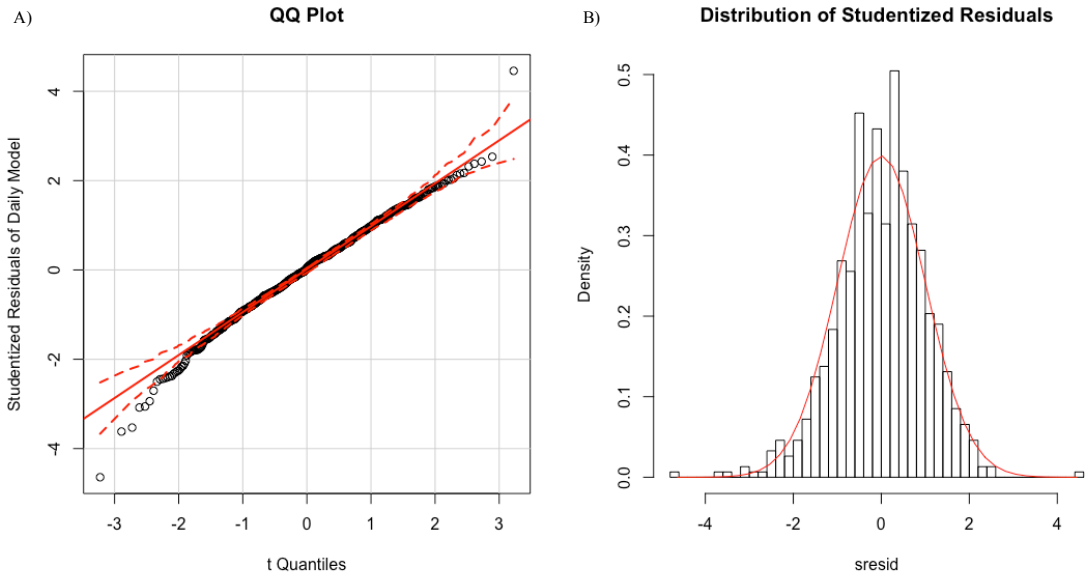


Figure 6- 9. A) QQplot of studentized residuals from the daily model against theoretical quantiles. B) Histogram of studentized residuals. The red line indicates a normal distribution, as calculated from the minimum and maximum studentized residuals.

6.3.5 Measures of accuracy

The 95% confidence interval of the mean predicted transformed BAM values is between 2.07 and 2.14 (equates to 7.93 μg/m^3 and 8.52 μg/m^3). In other words, there is a 95% probability that the interval we obtained contains a true value of BAM $PM_{2.5}$ at the specified setting. The prediction interval for the transformed BAM is between 1.74 and 2.47 (equates

to 5.72 μg/m^3 and 11.82 μg/m^3). These are displayed in Figure 6- 10. The SE for the model is 0.1839 on 755 degrees of freedom. Confidence intervals for the parameter estimates are shown in Table 6-6.

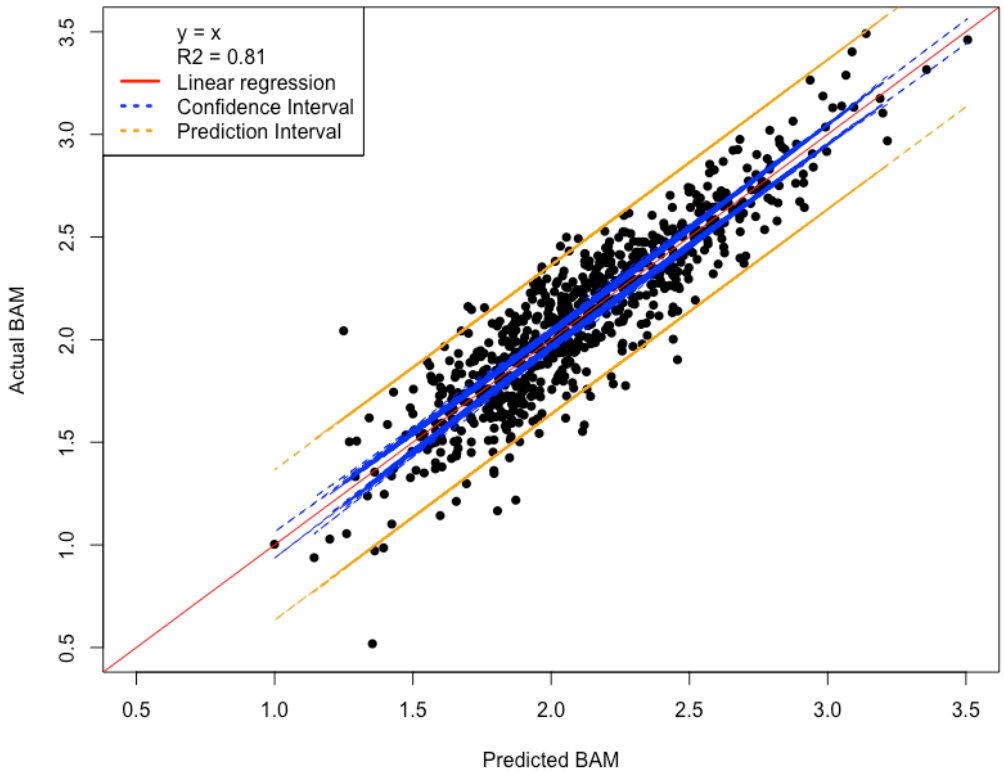


Figure 6- 10. Linear regression of actual and predicted BAM values over the collocated period. Confidence interval (blue), prediction interval (orange), linear regression (red) and R^2 value and coefficients are shown.

Table 6- 6. Parameter estimates and confidence intervals for the daily predictive model.

		Confidence interval	
Parameter	**Estimate**	**2.50%**	**97.50%**
(Intercept)	2.041	1.737	2.346
mthbkb	-0.050	-0.079	-0.021
rh	-0.005	-0.007	-0.003
nephl	0.395	0.340	0.449
nephl.l1	0.093	0.061	0.125
pm10l	0.232	0.175	0.291
lco	1.546	1.284	1.809
ws	0.044	0.021	0.068

6.3.6 Model validation and evaluation

One-step time-series cross validation was used to evaluate the models performance. This is the same procedure that was explained in Chapter 4 for the hourly prediction model. The model was developed on daily data from 03/09/2010 to 03/09/2011. The model was then applied on independent data, and re-defined each day, up until 29/11/2012.

The `modStats` function, from the *Openair* package was used to statistically evaluate the model, and the output is shown in Table 6- 7. The model was applied on 419 days of data. The FAC2 of 0.998 indicates that this model is very strong in its predictive ability. The MB is positive over all (0.019 $\mu g/m^3$), but seasonally this slightly varies. The MB is greatest in winter (-0.341 $\mu g/m^3$) and least in spring (0.185 $\mu g/m^3$). The spring modelled values possess the greatest spread from the observed values (MGE=1.446 $\mu g/m^3$). The correlations in all seasons are strong, with the highest Pearson's r in winter (r=0.951) and the lowest in summer (r=0.798). The IOA of autumn (0.812), winter (0.842) and spring (0.786) indicate a good performance of the model in these seasons. Same as for the hourly data, the COE is lowest in summer (0.350), indicating a poorer performance of the model for this season. The R^2 improves to 0.82.

Compared to the `modStats` output for the hourly model, it is clear through the evaluation statistics, the daily model performed better across all seasons than the hourly predictive model (Table 4- 5 compared to Table 6- 7).

Table 6- 7. Common numerical model evaluation statistics, based on predicted values from daily one-step time series cross validation.

Season	n	FAC2	MB ($\mu g/m^3$)	MGE ($\mu g/m^3$)	NMB	NMGE	RMSE	r	COE	IOA
Spring (SON)	168	1.000	0.185	1.446	0.019	0.152	1.998	0.903	0.572	0.786
Summer (DJF)	81	1.000	0.303	1.119	0.044	0.161	1.420	0.798	0.350	0.675
Autumn (MAM)	80	1.000	-0.215	1.152	-0.024	0.129	1.453	0.937	0.625	0.812
Winter (JJA)	90	0.989	-0.341	0.924	-0.039	0.105	1.184	0.951	0.684	0.842
All data	419	0.998	0.019	1.214	0.002	0.139	1.642	0.915	0.598	0.799

The time series plot of the daily predictions using time-series cross-validation show that the modelled values fit reasonably well with the actual recorded values (Figure 6- 11). Same as for the hourly model, the model does not capture extreme events all of the time. But the mean trend in the predicted values follows remarkably well.

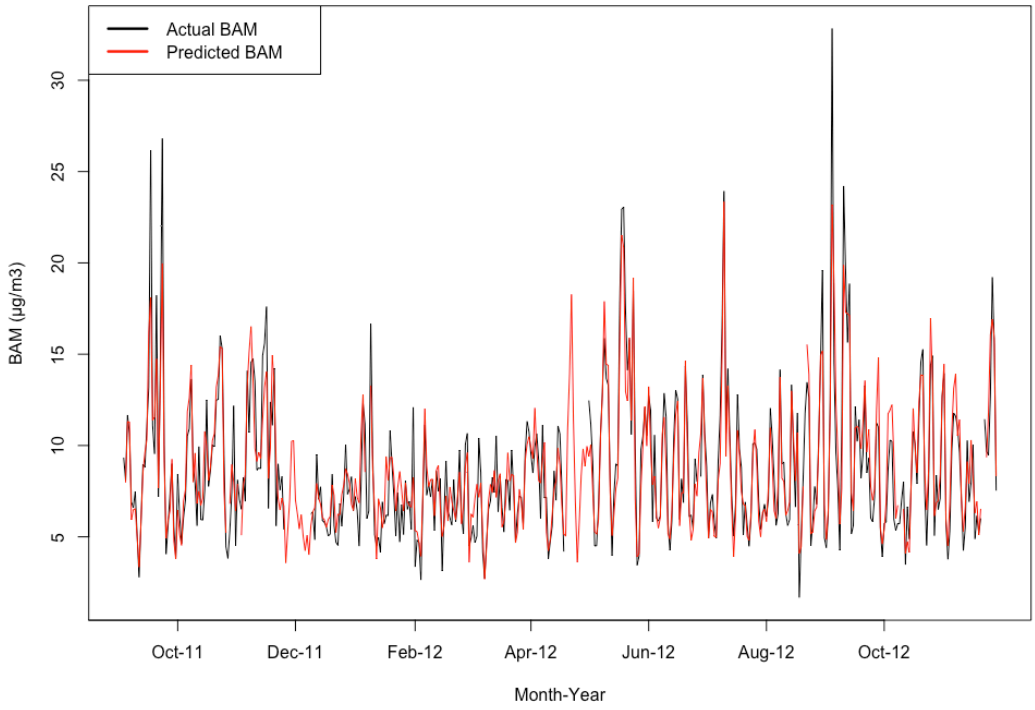


Figure 6- 11. Time series of actual BAM (black) and predicted BAM (red) values over the period of time when predictions were made using the time series cross validation, based on daily averages.

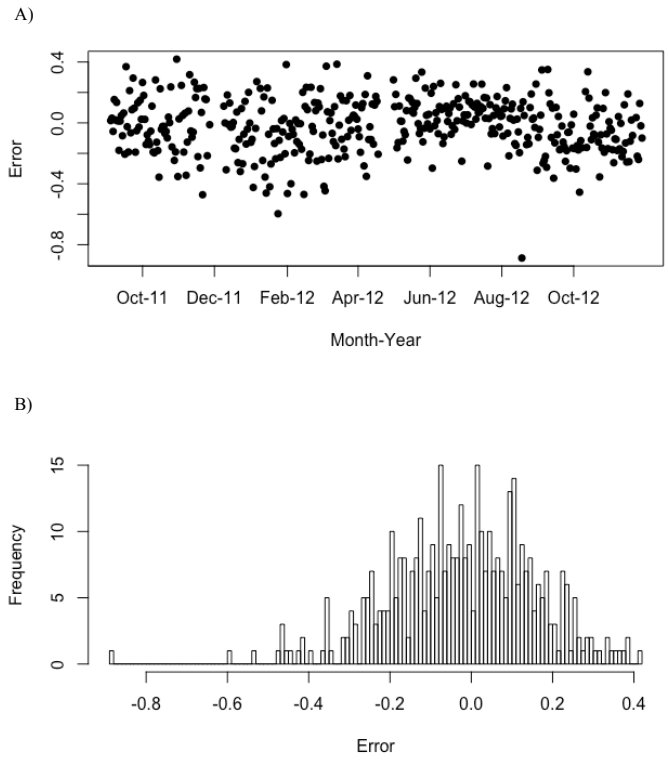


Figure 6- 12. Distribution of error over the period time where predictions were made using time-series cross-validation. A) showing a time series of the changes in error. B) showing a histogram of distribution of error.

An evaluation of the error, calculated as actual BAM (transformed) values minus predicted BAM (transformed) values, shows that there is still some scatter in the error terms (Figure 6- 12 A). From these 419 modelled values calculated using time-series cross validation, 46.8% of over predicted the actual values, and 53.2% under predicted actual values. The histogram and frequency suggests a normal Gaussian distribution of error terms (Figure 6- 12 B), meeting the assumption of normality for the model.

The predictive ability of the model is good on a daily basis (Figure 6- 13 A). Over the collocated period, the final model slightly under-predicts actual $PM_{2.5}$ values, with the greatest under-prediction occurring on Mondays and Tuesdays (by ~ 0.5 $\mu g/m^3$) (Figure 6- 13 A). All other days of the week the modelled values are remarkably close to the actual values (Figure 6- 13 A).

Additionally, the predictive ability of the final daily model is fantastic on a monthly basis (Figure 6- 13 B). The modelled values for August, September and December under-predict the actual values (Figure 6- 13 B). The under-prediction of these months is approximately 1.0 $\mu g/m^3$ for December, and approximately 0.5 $\mu g/m^3$ for August and September (Figure 6- 13 B). The remaining months track exceptionally well (Figure 6- 13 B). The modelled BAM values are a great improvement from the $PM_{2.5}$ TEOM values, providing a truer reading of actual $PM_{2.5}$ values on a daily and monthly basis (Figure 6- 13 A & B).

Looking at the yearly plot in Figure 6- 14, it is clear that the modelled values capture the seasonality of the TEOM quite well. The actual BAM readings are quite different from the summer of 2010/11 to 2011/12, and the modelled BAM values account for this change. Again, the modelled BAM values read a lot closer to the actual $PM_{2.5}$ BAM values than the $PM_{2.5}$ TEOM values do.

On a seasonal basis, autumn and spring model the actual $PM_{2.5}$ BAM values exceptionally well (Figure 6- 15). The winter over-predicts actual BAM values on Monday, and under-predicts on Wednesday, while summer under-predicts a lot more than any other season, especially on Monday, Tuesday and Friday (Figure 6- 15).

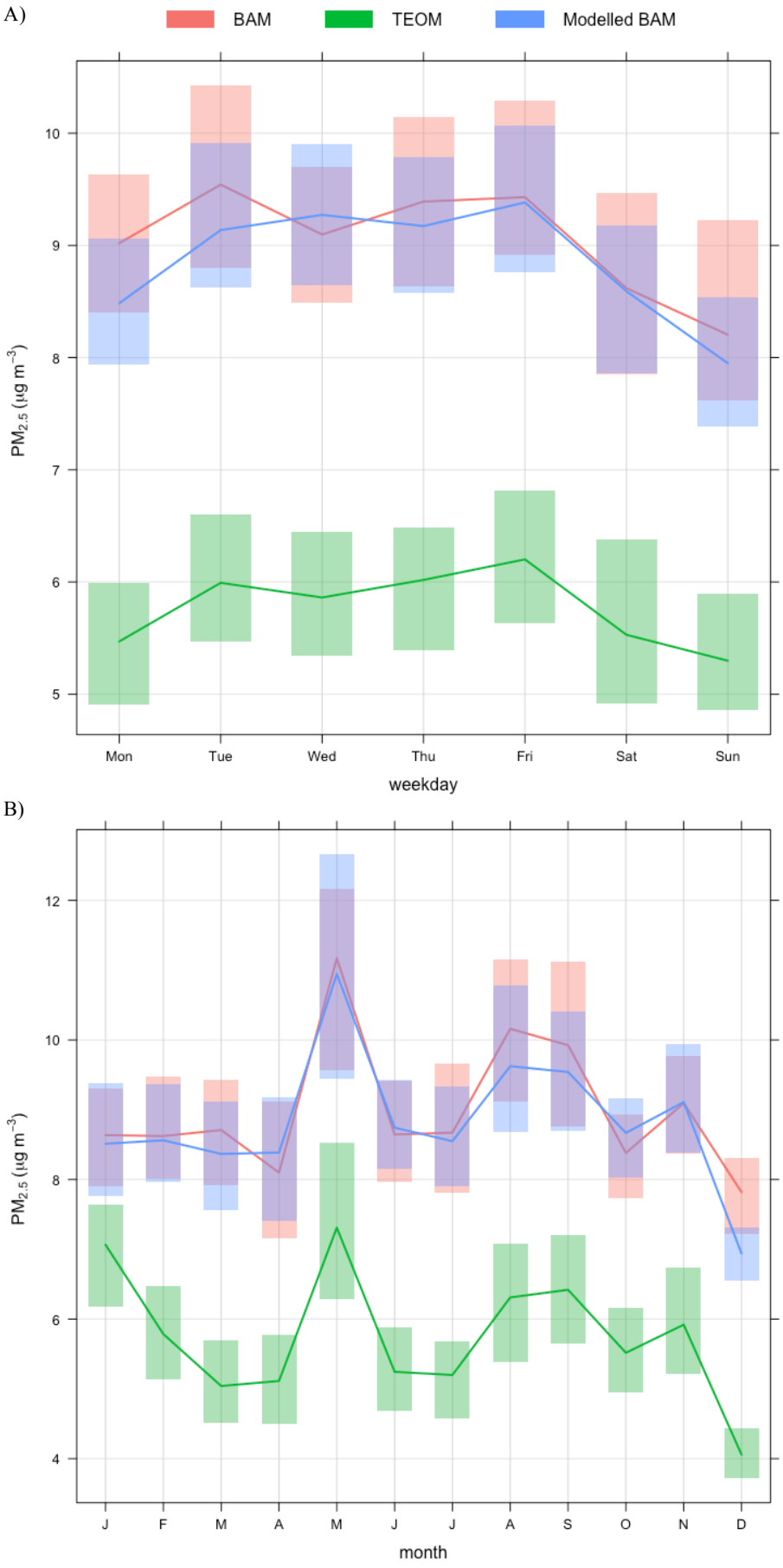


Figure 6- 13. Time Variation plot showing the original BAM (red) and TEOM (green) from the collocated period. The modelled BAM values are shown in blue. A) shows daily and B) shows monthly average plots. The shading around the boxes indicates a 95% confidence interval.

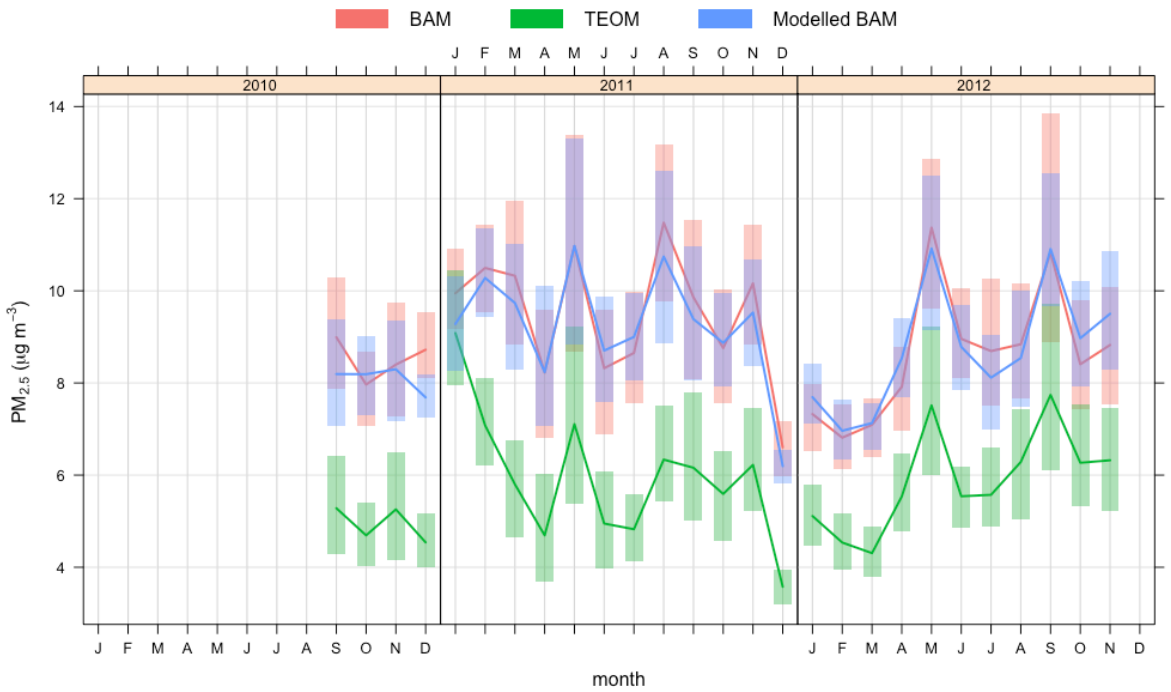


Figure 6- 14. Time variation plot showing the original BAM (red) and TEOM (green) from the collocated period, along with the modelled BAM values for the collocated period, shown by the blue line. The monthly averages are broken up by year.

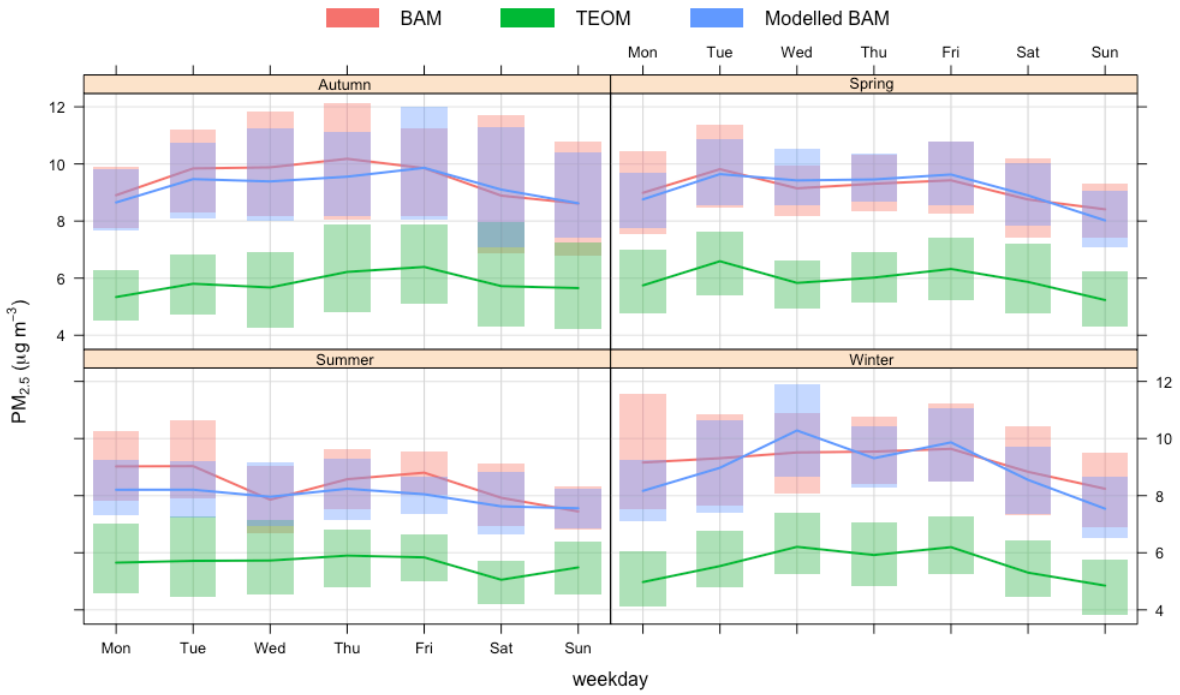


Figure 6- 15. Time variation plot showing the original BAM (red) and TEOM (green) from the collocated period, along with the modelled BAM values for the collocated period, shown by the blue line. The daily averages are broken up by season.

6.3.7 Ranking covariates by importance for prediction

Knowing which variables contribute the most to explaining the predictor variable is of prime importance. The change in R^2 value is recorded when the variable being analysed is added to the model as the last variable, helping to explain the unique variance that each covariate accounts for, beyond the other variables in the model. NEPH is ranked as the most important variable in the daily predictive model (R^2 difference = 0.0522) , followed by CO and PM_{10} (Table 6- 8).

Table 6- 8. Table showing each variable in the final daily model, the R^2 value when the particular variable was not included in model, the difference between the initial model (R^2 = 0.8059) and the model with that particular variable excluded. The variables were then ranked in terms of their importance, with the variable possessing the highest difference deemed having the highest importance.

Variable	R^2 when variable not included in model	Difference	Rank of importance
nephl	0.7537	0.0522	1
lco	0.7695	0.0364	2
pm10l	0.7929	0.0130	3
rh	0.7971	0.0088	4
nephl.l1	0.7976	0.0083	5
ws	0.8023	0.0036	6
Mthbk	0.8029	0.0030	7

6.3.8 ARDL model using only nephelometry data as a predictor

Given the high importance of the nephelometry parameter in terms of predicting BAM, and upon request from the OEH, a model using only NEPH values as an independent variable was constructed. The OEH saw fit for this type of model to be explored for two reasons; 1) to determine the necessity of including other variables for BAM prediction, making model application quicker and easier, and 2) to assess the predictive power of the NEPH alone. The results are outlined in Appendix 11. The results indicate that the daily model using NEPH as the only independent variable was not any better than what a daily model using the TEOM as the only independent variable could provide (Figure AP11-6 in Appendix 11). Whilst both the NEPH only and TEOM only model are ok, they do not perform as well as the daily predictive model with no limits on the input variables.

6.4 Application

In this section, the daily ARDL model is applied to the measurements at the Chullora site, from 24/01/2004 through to 28/11/2012. For this study the daily model was applied to values for which all covariates were available, a total of 2,925 observations over the ~9 year period.

The summary statistics in Table 6- 9 assist in assessing how the distribution of $PM_{2.5}$ in the period of 2004 to 2012 may have changed, using predictions from the daily ARDL model. 2004 has the highest mean modelled $PM_{2.5}$ value of 9.04 $\mu g/m^3$. From 2004 to 2008, modelled $PM_{2.5}$ values decrease over time, until 2009 where the modelled mean value rises to 8.78 $\mu g/m^3$, and remains fairly constant over the following years (Table 6- 9). Comparing the modelled values to the actual values, the mean fitted values do under-predict the mean actual values for 2011 and 2012 (2011 mean: modelled $PM_{2.5}$ = 8.89 $\mu g/m^3$, actual $PM_{2.5}$ = 9.56 $\mu g/m^3$; 2012 mean: modelled $PM_{2.5}$ = 8.54 $\mu g/m^3$, actual $PM_{2.5}$ = 8.67 $\mu g/m^3$), but not by much. The mean modelled $PM_{2.5}$ value for 2010 over predicts the mean actual $PM_{2.5}$ value, but this is likely a result of the actual 2010 results only having 117 observations, due to the installation of the BAM on 02/09/2010, and daily data only available from 03/09/2010 onwards for that year (Table 6- 9).

Table 6- 9. Summary statistics of modelled and actual BAM readings, for 2004 to 2012.

Year	Number of observations (n)	Mean ($\mu g/m^3$)	Standard deviation ($\mu g/m^3$)	Median ($\mu g/m^3$)	Min ($\mu g/m^3$)	Max ($\mu g/m^3$)	Standard error ($\mu g/m^3$)	NA's	% of data missing
Modelled values									
2004	274	9.04	4.24	8.29	1.66	28.48	0.26	69	20.12
2005	303	8.28	3.99	7.26	2.49	34.08	0.23	62	16.99
2006	331	8.52	3.56	7.84	2.89	27.32	0..20	34	9.32
2007	325	7.66	3.27	7.16	2.14	21.03	0.28	40	10.96
2008	333	7.28	2.75	6.92	2.39	18.45	0.25	33	9.02
2009	326	8.78	4.25	7.91	3.05	34.86	0.24	39	10.68
2010	352	8.72	3.84	7.88	2.95	25.64	0.20	12	3.30
2011	354	8.89	3.88	7.96	3.29	33.01	0.21	11	3.01
2012	327	8.54	3.52	7.71	2.64	23.73	0.19	7	2.10
Actual values									
2010	117	8.51	2.95	8.46	2.55	20.82	0.27	3	2.50
2011	346	9.56	4.43	8.65	2.80	31.86	0.24	19	5.21
2012	320	8.67	4.01	7.83	1.68	32.83	0.22	14	4.19

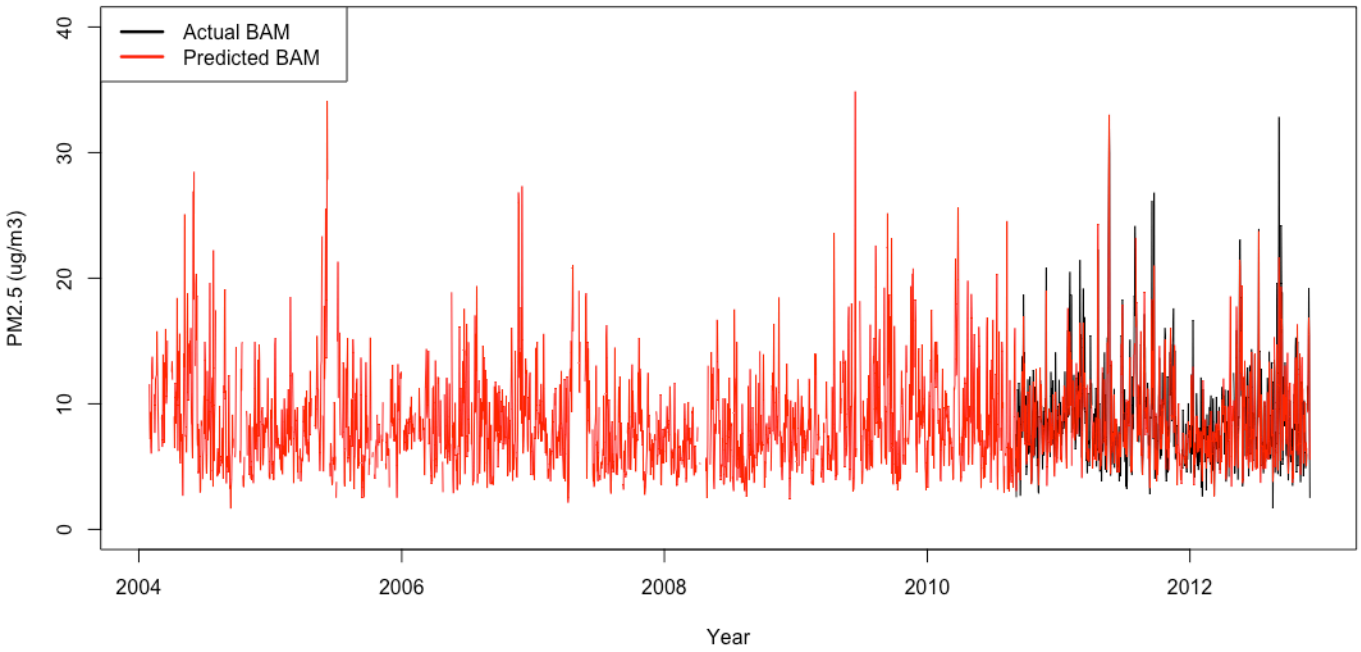


Figure 6- 16. Time series of daily data, showing the actual BAM (black) and the modelled BAM (red) for the period from 24/01/2004 to 29/11/2012.

The modelled (2004 to 2012) values demonstrate a 0.25% increase in $PM_{2.5}$ concentrations per year (95% confidence interval of -0.91%, 1.5%). However, the modelled (2004 to 2010) in combination with the actual (2010 to 2012) values, exhibit a 0.62 increase in $PM_{2.5}$ concentrations per year (95% confidence interval of -0.53%, 2.03%) (Figure 6- 17). Although this increase is not statistically significant, it further supports the idea that the daily modelled values do slightly under predict the actual $PM_{2.5}$ concentrations. There is a statistically significant increase in $PM_{2.5}$ concentrations in spring of 4.93% (95% confidence interval of 3.41%, 6.10%), between 2004 and 2012 (Figure 6- 18).

6.5 Summary

This daily model performs well in predicting daily $PM_{2.5}$ BAM values (Table 6- 7 and Figure 6- 13, Figure 6- 14 and Figure 6- 15). There is a strong relationship between the modelled and actual BAM readings for the model developed using all values over the collocated period ($R^2 = 0.81$). Therefore, the time series exhibited in Figure 6- 16, and the time variation plots with the model applied on data from 2004 to 2012 (Figure 6- 19), provide a great indication of what the daily average $PM_{2.5}$ level were from 2004 to 2012. The modelled (2004 to 2010) and actual (2010 to 2012) BAM values indicate an overall increase in $PM_{2.5}$ from 2004 to 2012 of 0.62% per year (-0.53%, 2.03%), although this increase is not

statistically significant. The increase of $PM_{2.5}$ in spring is statistically significant, increasing by 4.93% per year, (3.41%, 6.10%).

Additionally, the daily model is fit for application, as exceedances of ambient daily concentrations according to the NEPM Advisory Reporting Standards are based on daily averages, with the upper limit set to 25 $\mu g/m^3$ (in Table 1- 1) (Australian Government Department of Environment and Energy, 2014). According to our daily predictive model, there are 14 days between 24/01/2004 and 29/11/2012 that exceed this standard. These are exhibited in Table 6- 10.

Daily models containing a) only NEPH and b) only TEOM as independent variables for BAM predictions were constructed and evaluated. Both models have a satisfactory predictive ability, yet they are not as good as the daily model containing no limitations on the covariates. The summary statistics (Table AP11-1 in Appendix 11) and time variation plots (AP11-6 and AP11-7 in Appendix 11) suggest that the NEPH only model is no better than the TEOM only model. Ultimately, it becomes a trade-off for the user to decide the most appropriate model for their particular application, weighing up the complexity of the model against the its predictive ability.

Table 6- 10. Predicted $PM_{2.5}$ BAM values that exceed the standards set out in the Air NEPM (i.e. >25 $\mu g/m^3$).

Date	Predicted $PM_{2.5}$ BAM ($\mu g/m^3$)	Actual $PM_{2.5}$ BAM ($\mu g/m^3$)	Actual $PM_{2.5}$ TEOM ($\mu g/m^3$)
14/06/2009	34.86	NA	18.29
07/06/2005	34.08	NA	NA
21/05/2011	33.01	31.86	23.67
02/06/2004	28.48	NA	16.89
20/05/2011	28.11	27.55	19.38
08/06/2005	27.85	NA	25.53
01/12/2006	27.32	NA	31.95
31/05/2004	26.90	NA	19.50
21/11/2006	26.81	NA	31.13
22/11/2006	26.01	NA	NA
27/03/2010	25.64	NA	24.32
03/06/2005	25.50	NA	17.82
12/09/2009	25.16	NA	25.22
07/05/2004	25.08	NA	22.10

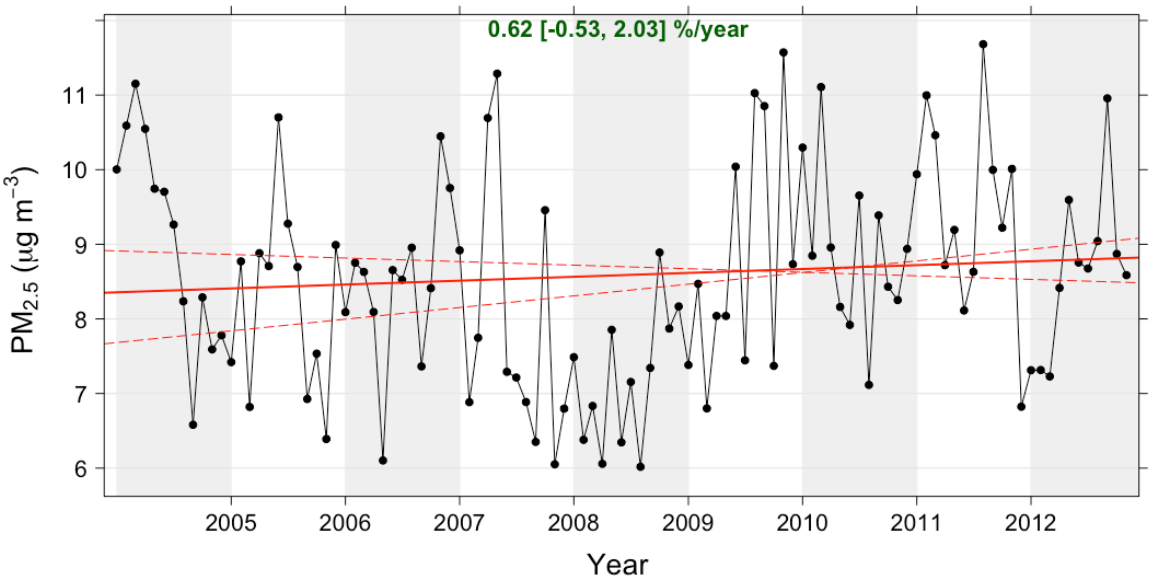


Figure 6- 17. Change in $PM_{2.5}$ from 2004 to 2012 based on the modelled (2004 to 2010) and actual values (2010 to 2012). Also shown is the average % increase in $PM_{2.5}$ per year with 95% confidence intervals. The increase is not statistically significant as there are no stars indicating significance next to the percent changes.

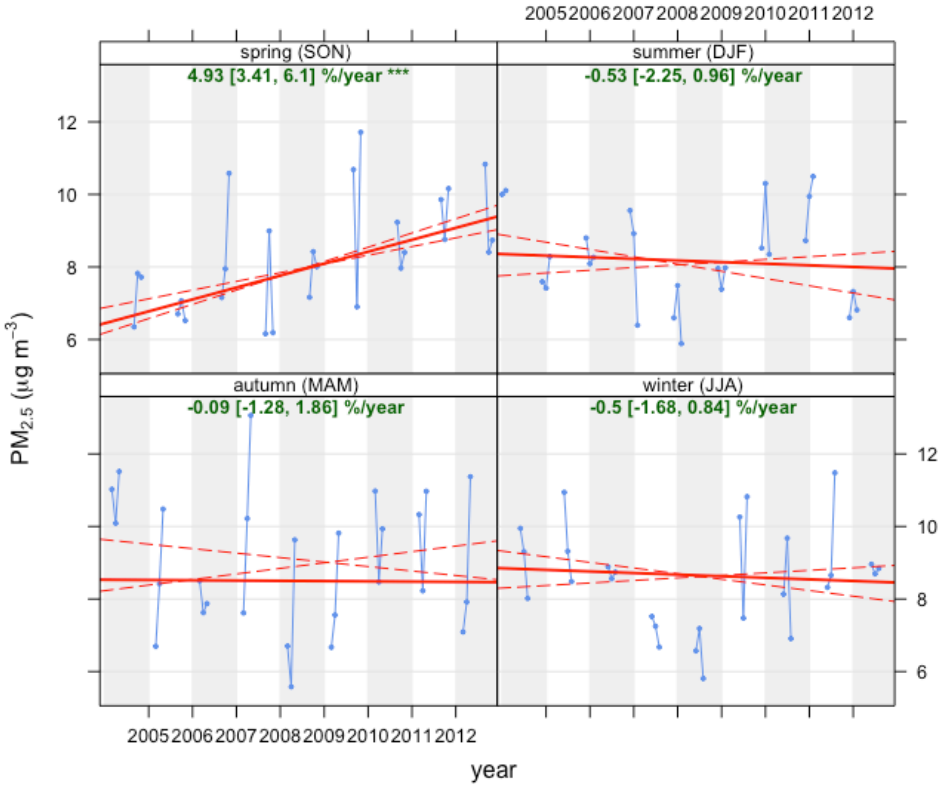


Figure 6- 18. Change in $PM_{2.5}$ from 2004 to 2012 shown seasonally, based on the modelled (2004 to 2010) and actual values (2010 to 2012). Also shown is the average % decrease or increase in $PM_{2.5}$ per year with 95% confidence intervals. The three green stars next to the % change in spring indicates the change in $PM_{2.5}$ over spring every year is statistically significant. The change in $PM_{2.5}$ in other seasons is not statistically significant.

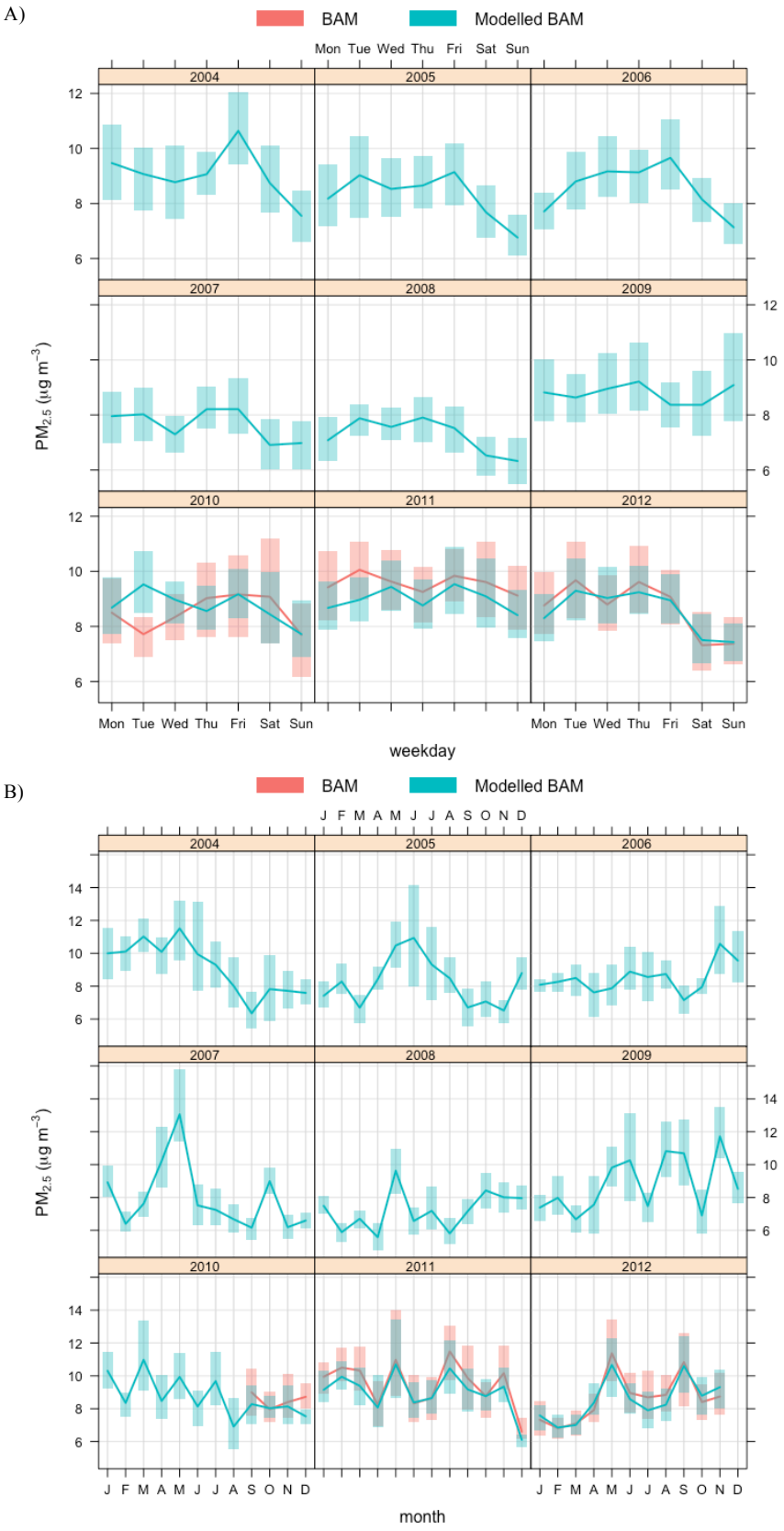


Figure 6- 19. Time Variation plots of the average daily actual BAM (red) and the modelled BAM (blue) readings from 2004 to 2012, with A) showing the average values for day of the week, and B) showing average values at a monthly scale.

Chapter 7: Discussion, conclusion and recommendations.

Discussion and conclusion

It is important to have a long consistent record of $PM_{2.5}$ to allow for long-term trend analysis. In this study, we developed four ARDL models to estimate $PM_{2.5}$ BAM concentrations at Chullora, Sydney, enabling the prediction of the ambient exposures for this site when actual BAM measurements were not available. Local meteorological, air pollution and gas covariates were integrated into a single model for $PM_{2.5}$ predictions, at hourly and daily intervals. The models captured linear relationships amongst the covariates and the $PM_{2.5}$ concentrations.

A linear model was appropriate for this study, improving the R^2 for hourly data (from 0.24 to 0.43) and daily data (from 0.75 to 0.81). A lot of studies use linear regression and correlation of collocated measurements to make different measurement methods comparable. Fu et al. (2014) corrected TEOM data to align with the collocated Federal Reference Method, by using linear regression, with 16 of the 23 sites possessing an $R^2 > 0.8$. Blanchard et al. (2011) used linear equations to standardize all of their data from samplers over California from 1980 - 2007, converting fine mass measurements from different methods to a standard Federal Reference Method value. Their study successfully reconstructed the historical $PM_{2.5}$ database with a high degree of accuracy ($R^2 > 0.9$, mean absolute error ~2.0 $\mu g/m^3$) using a range of covariates in a linear regression. Watson and Chow (2002) and Park et al. (2006) evaluated equivalence, comparability, and predictability between instruments using linear regression, and compared it to the requirements set out by United States Environmental Protection Agency (1997) regarding the equivalence of instruments. Whereas Clements et al. (2016) estimated the daily average semi-volatile fraction of $PM_{2.5}$ from the total $PM_{2.5}$ concentration using linear regression.

Many authors emphasise the importance of the construction of a model that is able to account for changes in meteorological conditions and emission sources, including influential variables, like temperature, relative humidity, and particle composition, rather than a statistical single correction factor (Charron et al., 2004, Green et al., 2001, Gehrig et al., 2005, Kashuba and Scheff, 2008). Our models do account for meteorological variations, yet they fail to account for the changes in particle composition.

While the statistical output of the hourly models performance is satisfactory (R^2 = 0.43, Pearson's correlation r = 0.80, IOA = 0.70, FAC2 = 0.81, MB = -0.43 $\mu g/m^3$ and RMSE

= 3.95, between the fitted and observed $PM_{2.5}$ BAM values tested on independent samples), it still has autocorrelation in its residuals suggesting there is still information left in the residuals that should be included when computing the forecast (Hyndman and Athanasopoulos, 2013). This research failed to develop a statistically robust hourly model, therefore this hourly model requires further refinement if it is to be used to correct hourly TEOM values.

Our daily predictive model produced a stable estimation of the time series, with a high R^2 between fitted and actual values (0.81). The statistical evaluation of the daily model, including Pearson's correlation (0.92), the FAC2 (1.00), IOA (0.80) MB (0.02 $\mu g/m^3$) and MGE (1.21 $\mu g/m^3$) between the predicted and observed concentrations based on independent samples, all indicate a well performing model. The model is statistically robust, and suitable for forecasting historical $PM_{2.5}$ concentrations for independent samples as demonstrated through the statistical output and time variation plots from the time-series cross validation (Table 6- 7, Figure 6- 5 and Figure 6- 10). It can be used to determine exceedances of $PM_{2.5}$ standards, and show how the $PM_{2.5}$ distributions have changed over time.

An Air Quality Index is a number used to communicate the overall level of pollution in a particular area, influenced by the population exposed. The OEH wish to back-calculate the Air Quality Index from 2004 to 2010, when the TEOM 1400AB was operational in New South Wales. The daily ARDL model is suitable to be applied to correct the TEOM measurements between 2004 and 2010 for Chullora, and therefore, is suitable to be used to assist in calculating the Air Quality Index for Chullora during this time.

Given the Advisory Reporting Standards for $PM_{2.5}$ by the NEPM are reported as daily averages (maximum ambient concentration of 25 $\mu g/m^3$over 1 day), the daily model can be used to determine exceedances of these ambient air $PM_{2.5}$ standards. According to the daily predictive model, there are 14 days from 2004 to 2012 where the $PM_{2.5}$ concentration exceeded the daily 25 $\mu g/m^3$ limit (see Table 6- 10).

Based on the modelled (2004 to 2010) and actual (2010 to 2012) daily values, there is a statistically significant increase in $PM_{2.5}$ concentrations in spring from 2004 to 2012 (4.93% per year with a 95% confidence interval of 3.41%, 6.10%). However, based on all of the seasons combined, the results suggest an increase in $PM_{2.5}$ from 2004 to 2012, but this is not a statistically significant increase (0.62% per year with a 95% confidence interval of -0.53%, 2.03%).

In this study, nephelometry data was determined as the most important factor in determining daily $PM_{2.5}$ concentrations for Chullora (R^2 difference = 0.05). For this reason,

and upon request from the OEH, a daily model using only nephelometry data as the predictor variable was constructed to determine its predictive ability. The models performance was satisfactory, (Pearson's correlation r = 0.86, COE = 0.51, IOA = 0.75, MB = 0.150 μg/m^3, MGE = 1.48 μg/m^3 and RMSE = 2.09 between predicted and observed $PM_{2.5}$ values based on independent samples), although it was determined that a model using only the TEOM data as the predictor variable performed just as well (Pearson's correlation r = 0.91, COE = 0.54, IOA = 0.77, MB = 0.24 μg/m^3, MGE = 1.38 μg/m^3 and RMSE = 1.86 between predicted and observed $PM_{2.5}$ values based on independent samples). However, both of these were not as good as the daily model with no limitation on input variables. The addition of more variables improves the statistical performance of the model. Therefore, when choosing to apply the daily model with or without a restriction on the covariates, it becomes a tradeoff for the user to decide between model simplicity and improved model performance.

Recommendations for future research

Drawbacks in our ARDL models invite future research. Further studies should consider the inclusion of $PM_{2.5}$ BAM data from an additional site that is in close proximity to Chullora. $PM_{2.5}$ BAM data from another site can be adjusted and correlated with the Chullora site, then incorporated into a predictive model. This would be useful as we have demonstrated that the autocorrelation of the hourly model improves significantly when previous readings of BAM (BAM lags) are included in the predictive model (Figure 4- 4).

Additionally, future studies should incorporate particle composition data sourced from ANSTO's Aerosol Sampling Program. Relevant literature has demonstrated that accounting for particle composition in the modelling of $PM_{2.5}$ can improve the agreement between instruments, as the semi volatile material lost could potentially be accounted for (Lee et al., 2005, Hauck et al., 2004, Godri et al., 2009, Li et al., 2012, Schwab et al., 2006, Chung et al., 2001, Clements, 2013). ANSTO (2010) and Cohen et al. (2016) confirm the significant contribution that ammonium sulfate makes to $PM_{2.5}$ composition in Sydney, highlighting its variability on a seasonal basis. Capturing these seasonal changes in a model may lead to a better performing model, especially in the summer season. Using this type of information, or other fundamental gaseous precursors to PM like volatile organic compounds, may improve the performance of the hourly and daily model, and may assist in reducing the autocorrelation in the residuals of the hourly model.

Researching, testing and applying an appropriate imputation method for the covariates to maximize the number of predictions made should be considered in future research, as

missing data can cause bias as a result of the systematic differences amongst the observed and unobserved data (Norazian et al., 2013, Norazian et al., 2008), while reducing the sample size and power of study (Allison, 2002). This would improve the confidence we could have in our results, and the conclusions drawn from these results.

Additionally, advancing the statistical techniques used in the development of the model should be considered for future research. Given the nature of the data, we acknowledge that the covariates themselves are not free from error. Therefore, error-in-variable approaches like Orthogonal and Deming regression should be explored to see if they provide a better model (Deming, 1943, Bilonick et al., 2015). Orthogonal regression equations were used by Hsu et al. (2016) to adjust TEOM readings to improve its agreement with the FRM, reducing the relative difference from 18% to 13% during cold seasons. Alternatively, while considerably more complex, a states based modelling approach, such as Kalman filter, could be employed into future research to serve as a framework for powerful space time modelling of these atmospheric processes (Ghil and Malanotte-Rizzoli, 1991, as reported in Heemink and Segers, 2002).

Non-linear regression could be investigated for its suitability to be applied to this data, with literature revealing that non-linear regression can out-perform linear regression (Li et al., 2017, Bilger and Manning, 2011, Li et al., 2013). Artificial neural networks may be a suitable non-linear tool for pollution forecasting, using multilayer perceptron architecture (Díaz-Robles et al., 2008, Thomas and Jacko, 2007, Sofuoglu et al., 2006). However, non-linear regression can cause over-fitting that may cause bias in its predictions (Li et al., 2013), with its application being a trade-off between the simplicity of the model and the extent of its statistical suitability (Kashuba and Scheff, 2008).

Additionally, it is important to be able to predict $PM_{2.5}$ at a variety of locations, other than just the one study site. These hourly and daily predictive models are limited in their geographical extent, by the fact that they have only been tested on independent samples for the Chullora site. Therefore, they are limited in their application, to the Chullora site only. Future research should encompass applying this model beyond Chullora, where independent samples can be tested, and the model evaluated. Once completed, this will assist in back-calculating the Air Quality Index between 2004 and 2012 beyond Chullora.

Chapter 8: References

AGUINIS, H., GOTTFREDSON, R. K. & JOO, H. 2013. Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods,* 16**,** 270-301.

ALLEN, G., SIOUTAS, C., KOUTRAKIS, P., REISS, R., LURMANN, F. W. & ROBERTS, P. T. 1997. Evaluation of the TEOM® method for measurement of ambient particulate mass in urban areas. *Journal of the Air & Waste Management Association,* 47**,** 682-689.

ALLISON, P. D. 2002. Missing data: Quantitative applications in the social sciences. *British Journal of Mathematical and Statistical Psychology,* 55**,** 193-196.

ANSTO 2010. Fine particulate aerosol sampling newsletter. Lucas Heights, NSW: Australian Nuclear Science and Technology Organisation.

ARLOT, S. & CELISSE, A. 2010. A survey of cross-validation procedures for model selection. *Statistics surveys,* 4**,** 40-79.

AUSTRALIAN GOVERNMENT DEPARTMENT OF ENVIRONMENT AND ENERGY 2014. Draft Variation to the National Environment Protection (Ambient Air Quality) Measure – Impact Statement. Canberra.

AYERS, G., KEYWOOD, M. & GRAS, J. 1999. TEOM vs. manual gravimetric methods for determination of PM2. 5 aerosol mass concentrations. *Atmospheric Environment,* 33**,** 3717-3721.

BARNETT, A. G., WILLIAMS, G. M., SCHWARTZ, J., BEST, T. L., NELLER, A. H., PETROESCHEVSKY, A. L. & SIMPSON, R. W. 2006. The effects of air pollution on hospitalizations for cardiovascular disease in elderly people in Australian and New Zealand cities. *Environmental health perspectives***,** 1018-1023.

BARNETT, A. G., WILLIAMS, G. M., SCHWARTZ, J., NELLER, A. H., BEST, T. L., PETROESCHEVSKY, A. L. & SIMPSON, R. W. 2005. Air pollution and child respiratory health: a case-crossover study in Australia and New Zealand. *American journal of respiratory and critical care medicine,* 171**,** 1272-1278.

BEAVER, S., PALAZOGLU, A., SINGH, A., SOONG, S.-T. & TANRIKULU, S. 2010. Identification of weather patterns impacting 24-h average fine particulate matter pollution. *Atmospheric Environment,* 44**,** 1761-1771.

BELL, M. L., EBISU, K., PENG, R. D., SAMET, J. M. & DOMINICI, F. 2009. Hospital admissions and chemical composition of fine particle air pollution. *American journal of respiratory and critical care medicine,* 179**,** 1115-1120.

BILGER, M. & MANNING, W. G. 2011. Measuring overfitting and mispecification in nonlinear models. *Health, Econometrics and Data Group Working Paper,* 11**,** 25.

BILONICK, R. A., CONNELL, D. P., TALBOTT, E. O., RAGER, J. R. & XUE, T. 2015. Using structural equation modeling to construct calibration equations relating PM 2.5 mass concentration samplers to the federal reference method sampler. *Atmospheric Environment,* 103**,** 365-377.

BLANCHARD, C. L., TANENBAUM, S. & MOTALLEBI, N. 2011. Spatial and Temporal Characterization of PM2. 5 Mass Concentrations in California, 1980–2007. *Journal of the Air & Waste Management Association,* 61**,** 339-351.

BOUIS, P. A. 1999. *Air Pollution,* Unites States of America, CRC Press.

BROOK, R. D., RAJAGOPALAN, S., POPE, C. A., BROOK, J. R., BHATNAGAR, A., DIEZ-ROUX, A. V., HOLGUIN, F., HONG, Y., LUEPKER, R. V. & MITTLEMAN, M. A. 2010. Particulate matter air pollution and cardiovascular disease. *Circulation,* 121**,** 2331-2378.

BROWN, D. M., WILSON, M. R., MACNEE, W., STONE, V. & DONALDSON, K. 2001. Size-dependent proinflammatory effects of ultrafine polystyrene particles: a role for surface area and oxidative stress in the enhanced activity of ultrafines. *Toxicology and applied pharmacology,* 175**,** 191-199.

CARSLAW, D. C. & ROPKINS, K. 2012. Openair—an R package for air quality data analysis. *Environmental Modelling & Software,* 27**,** 52-61.

CHAN, Y.-C., COHEN, D. D., HAWAS, O., STELCER, E., SIMPSON, R., DENISON, L., WONG, N., HODGE, M., COMINO, E. & CARSWELL, S. 2008. Apportionment of sources of fine and coarse particles in four major Australian cities by positive matrix factorisation. *Atmospheric Environment,* 42**,** 374-389.

CHARRON, A., HARRISON, R. M., MOORCROFT, S. & BOOKER, J. 2004. Quantitative interpretation of divergence between PM 10 and PM 2.5 mass measurement by TEOM and gravimetric (Partisol) instruments. *Atmospheric Environment,* 38**,** 415-423.

CHOW, J. C. 1995. Measurement methods to determine compliance with ambient air quality standards for suspended particles. *Journal of the Air & Waste Management Association,* 45**,** 320-382.

CHUNG, A., CHANG, D. P., KLEEMAN, M. J., PERRY, K. D., CAHILL, T. A., DUTCHER, D., MCDOUGALL, E. M. & STROUD, K. 2001. Comparison of real-time instruments used to monitor airborne particulate matter. *Journal of the Air & Waste Management Association,* 51**,** 109-120.

CLEMENTS, N., HANNIGAN, M. P., MILLER, S. L., PEEL, J. L. & MILFORD, J. B. 2016. Comparisons of urban and rural PM 10− 2.5 and PM 2.5 mass concentrations and semi-volatile fractions in northeastern Colorado. *Atmospheric Chemistry and Physics,* 16**,** 7469-7484.

CLEMENTS, N. S. 2013. *The CCRUSH study: Characterization of coarse and fine particulate matter in northeastern Colorado.* University of Colorado at Boulder.

COHEN, D., ATANACIO, A., STELCER, E. & GARTON, D. 2016. Sydney Particle Characterisation Study: PM2.5 Source Appointment in the Sydney Region between 2000 and 2014. ANSTO.

COHEN, D. D., CRAWFORD, J., STELCER, E. & ATANACIO, A. J. 2012. Application of positive matrix factorization, multi-linear engine and back trajectory techniques to the quantification of coal-fired power station pollution in metropolitan Sydney. *Atmospheric environment,* 61**,** 204-211.

COHEN, D. D., STELCER, E., GARTON, D. & CRAWFORD, J. 2011. Fine particle characterisation, source apportionment and long-range dust transport into the Sydney Basin: a long term study between 1998 and 2009. *Atmospheric Pollution Research,* 2**,** 182-189.

COPE, M., KEYWOOD, M., EMMERSON, K., GALBALLY, I., BOAST, K., CHAMBERS, S., CHENG, M., CRUMEYROLLE, S., DUNNE, E., FEDELE, R., GILLETT, R., GRIFFITHS, A., HARNWELL, J., KATZFEY, J., HESS, D., LAWSON, S., MILJEVIC, B., MOLLOY, S., POWELL, J., REISEN, F., RISTOVSKI, Z., SELLECK, P., WARD, J., ZHANG, C. & ZENG, J. 2014. Sydney Particle Study - Stage-II.

CORTINA, J. M. 2002. Big things have small beginnings: An assortment of “minor” methodological misunderstandings. *Journal of Management,* 28**,** 339-362.

CRAWFORD, J., CHAMBERS, S., COHEN, D., WILLIAMS, A., GRIFFITHS, A. & STELCER, E. 2016a. Assessing the impact of atmospheric stability on locally and remotely sourced

aerosols at Richmond, Australia, using Radon-222. *Atmospheric Environment,* 127**,** 107-117.

CRAWFORD, J., GRIFFITHS, A., COHEN, D. D., JIANG, N. & STELCER, E. 2016b. Particulate pollution in the Sydney region: source diagnostics and synoptic controls. *Aerosol and Air Quality Research,* 16.

CYRYS, J., DIETRICH, G., KREYLING, W., TUCH, T. & HEINRICH, J. 2001. PM 2.5 measurements in ambient aerosol: comparison between Harvard impactor (HI) and the tapered element oscillating microbalance (TEOM) system. *Science of the total environment,* 278**,** 191-197.

DAVIS, R. E. & GAY, D. A. 1993. An assessment of air quality variations in the south-western USA using an upper air synoptic climatology. *International Journal of Climatology,* 13**,** 755-781.

DAYAN, U. & LEVY, I. 2005. The influence of meteorological conditions and atmospheric circulation types on PM10 and visibility in Tel Aviv. *Journal of Applied Meteorology,* 44**,** 606-619.

DEMING, W. E. 1943. Statistical adjustment of data.

DÍAZ-ROBLES, L. A., ORTEGA, J. C., FU, J. S., REED, G. D., CHOW, J. C., WATSON, J. G. & MONCADA-HERRERA, J. A. 2008. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. *Atmospheric Environment,* 42**,** 8331-8340.

EATOUGH, D. J., EATOUGH, N. L., OBEIDI, F., PANG, Y., MODEY, W. & LONG, R. 2001. Continuous determination of PM2. 5 mass, including semi-volatile species. *Aerosol Science & Technology,* 34**,** 1-8.

EATOUGH, D. J., LONG, R. W., MODEY, W. K. & EATOUGH, N. L. 2003. Semi-volatile secondary organic aerosol in urban atmospheres: meeting a measurement challenge. *Atmospheric Environment,* 37**,** 1277-1292.

EKSTRÖM, M., MCTAINSH, G. H. & CHAPPELL, A. 2004. Australian dust storms: temporal trends and relationships with synoptic pressure distributions (1960–99). *International Journal of Climatology,* 24**,** 1581-1599.

FEDERAL REGISTER OF LEGISLATIVE INSTRUMENTS 2016. National Environment Protection (Ambient Air Quality Measure).

FERIN, J., OBERDÖRSTER, G., SODERHOLM, S. C. & GELEIN, R. 1991. Pulmonary tissue access of ultrafine particles. *Journal of aerosol medicine,* 4**,** 57-68.

FU, L., NUNIFU, T. & LEUNG, B. 2014. A two-step approach for relating tapered element oscillating microbalance and dichotomous air sampler PM2. 5 measurements. *Journal of the Air & Waste Management Association,* 64**,** 1195-1203.

GEHRIG, R., HUEGLIN, C., SCHWARZENBACH, B., SEITZ, T. & BUCHMANN, B. 2005. A new method to link PM10 concentrations from automatic monitors to the manual gravimetric reference method according to EN12341. *Atmospheric Environment,* 39**,** 2213-2223.

GHIL, M. & MALANOTTE-RIZZOLI, P. 1991. Data assimilation in meteorology and oceanography. *Advances in geophysics,* 33**,** 141-266.

GODRI, K., EVANS, G., SLOWIK, J., KNOX, A., ABBATT, J., BROOK, J., DANN, T. & DABEK-ZLOTORZYNSKA, E. 2009. Evaluation and application of a semi-continuous chemical characterization system for water soluble inorganic PM 2.5 and associated precursor gases. *Atmospheric Measurement Techniques,* 2**,** 65-80.

GREEN, D., FULLER, G. & BARRATT, B. 2001. Evaluation of TEOM TM 'correction factors' for assessing the EU Stage 1 limit values for PM10. *Atmospheric Environment,* 35**,** 2589-2593.

GREENE, D. S. 2005. Comparison Between Tapered Element Microbalance (TEOM) and Federal Reference Method (FRM) for PM 2.5 Measurement in East Tennessee.

GROVER, B. D. 2006. Measurement, Characterization, and Source Apportionment of the Major Chemical Components of Fine Particulate Material, Including Semi-Volatile Species.

GROVER, B. D., KLEINMAN, M., EATOUGH, N. L., EATOUGH, D. J., HOPKE, P. K., LONG, R. W., WILSON, W. E., MEYER, M. B. & AMBS, J. L. 2005. Measurement of total PM2. 5 mass (nonvolatile plus semivolatile) with the Filter Dynamic Measurement System tapered element oscillating microbalance monitor. *Journal of Geophysical Research: Atmospheres,* 110.

HAIKERWAL, A., AKRAM, M., DEL MONACO, A., SMITH, K., SIM, M. R., MEYER, M., TONKIN, A. M., ABRAMSON, M. J. & DENNEKAMP, M. 2015. Impact of fine particulate matter (PM2. 5) exposure during wildfires on cardiovascular health outcomes. *Journal of the American Heart Association,* 4**,** e001653.

HANSEN, A., BI, P. & NITSCHKE, M. 2009. Air pollution and cardiorespiratory health in Australia: the impact of climate change. *Environmental Health,* 9**,** 17.

HANSEN, A., BI, P., NITSCHKE, M., PISANIELLO, D., RYAN, P., SULLIVAN, T. & BARNETT, A. G. 2012. Particulate air pollution and cardiorespiratory hospital admissions in a temperate Australian city: a case-crossover analysis. *Science of the total environment,* 416**,** 48-52.

HAUCK, H., BERNER, A., GOMISCEK, B., STOPPER, S., PUXBAUM, H., KUNDI, M. & PREINING, O. 2004. On the equivalence of gravimetric PM data with TEOM and beta-attenuation measurements. *Journal of Aerosol Science,* 35**,** 1135-1149.

HAWTHORNE, G. & ELLIOTT, P. 2005. Imputing cross-sectional missing data: Comparison of common techniques. *Australian and New Zealand Journal of Psychiatry,* 39**,** 583-590.

HEEMINK, A. & SEGERS, A. 2002. Modeling and prediction of environmental data in space and time using Kalman filtering. *Stochastic Environmental Research and Risk Assessment,* 16**,** 225-240.

HSU, Y.-M., WANG, X., CHOW, J. C., WATSON, J. G. & PERCY, K. E. 2016. Collocated comparisons of continuous and filter-based PM2. 5 measurements at Fort McMurray, Alberta, Canada. *Journal of the Air & Waste Management Association,* 66**,** 329-339.

HUANG, C.-H. & TAI, C.-Y. 2008. Relative humidity effect on PM2. 5 readings recorded by collocated beta attenuation monitors. *Environmental Engineering Science,* 25**,** 1079-1090.

HUANG, X.-F., YU, J. Z., YUAN, Z., LAU, A. K. & LOUIE, P. K. 2009. Source analysis of high particulate matter days in Hong Kong. *Atmospheric environment,* 43**,** 1196-1203.

HUSAR, R. B. 1974. Atmospheric particulate mass monitoring with a β radiation detector. *Atmospheric Environment (1967),* 8**,** 183-188.

HYNDMAN, R. J. & ATHANASOPOULOS, G. 2013. Forecasting: Principles and Practice.

HYNDMAN, R. J. & KOEHLER, A. 2014. Measuring forecast accuracy. *Prieiga per internetą:<* http://www/. *robjhyndman. com/papers/forecast-accuracy. pdf>,[žiūrėta 2016 05 08]*.

JACOB, D. J., CRAWFORD, J. H., KLEB, M. M., CONNORS, V. S., BENDURA, R. J., RAPER, J. L., SACHSE, G. W., GILLE, J. C., EMMONS, L. & HEALD, C. L. 2003. Transport and Chemical Evolution over the Pacific (TRACE-P) aircraft mission: Design, execution, and first results. *Journal of Geophysical Research: Atmospheres,* 108.

JAFFE, D., MCKENDRY, I., ANDERSON, T. & PRICE, H. 2003. Six 'new'episodes of trans-Pacific transport of air pollutants. *Atmospheric Environment,* 37**,** 391-404.

JIANG, N., HAY, J. E. & FISHER, G. W. 2005. Synoptic weather types and morning rush hour nitrogen oxides concentrations during Auckland winters.

KAHN, H. D. 1973. Note on the distribution of air pollutants. *Journal of the Air Pollution Control Association,* 23**,** 973-973.

KAM, W. 2012. *Particulate matter (PM) exposure for commuters in Los Angeles: Chemical characterization and implications to public health*, University of Southern California.

KAN, H., LONDON, S. J., CHEN, G., ZHANG, Y., SONG, G., ZHAO, N., JIANG, L. & CHEN, B. 2007. Differentiating the effects of fine and coarse particles on daily mortality in Shanghai, China. *Environment international,* 33**,** 376-384.

KASHUBA, R. & SCHEFF, P. A. 2008. Nonlinear regression adjustments of multiple continuous monitoring methods produce effective characterization of short-term fine particulate matter. *Journal of the Air & Waste Management Association,* 58**,** 812-820.

LEE, J. H., HOPKE, P. K., HOLSEN, T. M. & POLISSAR, A. V. 2005. Evaluation of continuous and filter-based methods for measuring PM2. 5 mass concentration. *Aerosol science and technology,* 39**,** 290-303.

LEGATES, D. R. & MCCABE, G. J. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water resources research,* 35**,** 233-241.

LEGATES, D. R. & MCCABE, G. J. 2013. A refined index of model performance: a rejoinder. *International Journal of Climatology,* 33**,** 1053-1056.

LESLIE, L. M. & SPEER, M. S. 2006. Modelling dust transport over central eastern Australia. *Meteorological Applications,* 13**,** 141-167.

LEWTAS, J., PANG, Y., BOOTH, D., REIMER, S., EATOUGH, D. J. & GUNDEL, L. A. 2001. Comparison of sampling methods for semi-volatile organic carbon associated with PM2. 5. *Aerosol Science & Technology,* 34**,** 9-22.

LI, L., WU, A. H., CHENG, I., CHEN, J.-C. & WU, J. 2017. Spatiotemporal estimation of historical PM 2.5 concentrations using PM 10, meteorological variables, and spatial effect. *Atmospheric Environment*.

LI, L., WU, J., HUDDA, N., SIOUTAS, C., FRUIN, S. A. & DELFINO, R. J. 2013. Modeling the concentrations of on-road air pollutants in southern California. *Environmental science & technology,* 47**,** 9291-9299.

LI, Q.-F., WANG-LI, L., LIU, Z. & HEBER, A. J. 2012. Field evaluation of particulate matter measurements using tapered element oscillating microbalance in a layer house. *Journal of the Air & Waste Management Association,* 62**,** 322-335.

LILIENFELD, P. 1970. Beta-absorption-impactor aerosol mass monitor. *The American Industrial Hygiene Association Journal,* 31**,** 722-729.

LIM, S. S., VOS, T., FLAXMAN, A. D., DANAEI, G., SHIBUYA, K., ADAIR-ROHANI, H., ALMAZROA, M. A., AMANN, M., ANDERSON, H. R. & ANDREWS, K. G. 2013. A comparative risk assessment of burden of disease and injury attributable to 67 risk

factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The lancet,* 380**,** 2224-2260.
LIU, J. & MAUZERALL, D. L. 2005. Estimating the average time for inter-continental transport of air pollutants. *Geophysical research letters,* 32.
LONG, R. W., EATOUGH, N. L., MANGELSON, N. F., THOMPSON, W., FIET, K., SMITH, S., SMITH, R., EATOUGH, D. J., POPE, C. A. & WILSON, W. E. 2003. The measurement of PM 2.5, including semi-volatile components, in the EMPACT program: results from the Salt Lake City Study. *Atmospheric Environment,* 37**,** 4407-4417.
LONG, R. W., SMITH, R., SMITH, S., EATOUGH, N. L., MANGELSON, N. F., EATOUGH, D. J., POPE, C. A. & WILSON, W. E. 2002. Sources of fine particulate material along the Wasatch Front. *Energy & fuels,* 16**,** 282-293.
LUMLEY, T., DIEHR, P., EMERSON, S. & CHEN, L. 2002. The importance of the normality assumption in large public health data sets. *Annual review of public health,* 23**,** 151-169.
MACNEE, W. & DONALDSON, K. 2003. Mechanism of lung injury caused by PM10 and ultrafine particles with special reference to COPD. *European Respiratory Journal,* 21**,** 47s-51s.
MALM, W. C. 2000. Spatial and seasonal patterns and temporal variability of haze and its constituents in the United States.
MITCHELL, R., CAMPBELL, S. & QIN, Y. 2010. Recent increase in aerosol loading over the Australian arid zone. *Atmospheric Chemistry and Physics,* 10**,** 1689-1699.
MORGAN, G., BROOME, R. & JALALUDIN, B. 2013. Summary for Policy Makers of the Health Risk Assessment on Air Pollution in Australia. University Centre for Rural Health.
MUKAKA, M. M. 2012. A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal,* 24**,** 69-71.
MUSICK, D. 1999. A summary of the ambient air program for PM2.5. *Journal of Environmental Management***,** 17-22.
NATIONAL ENVIRONMENTAL PROTECTION COUNCIL. 2015. *National Environment Protection (Ambient Air Quality) Measure* [Online]. Available: http://www.nepc.gov.au/nepms/ambient-air-quality [Accessed 05/07/2017].
NATIONAL RESEARCH COUNCIL 1998. *Research Priorites for Airborne Particulate Matter: Volume 1: Immediate Priorites and a Long-Range Research Portfolio,* Washington, D.C., National Academy Press.
NATIONAL RESEARCH COUNCIL 2004. Research Priorities for Airborne Particulate Matter. Washington, D.C.
NELSON, A. C. A. J., T.R, 1980. Validation of Air Monitoring Data. United States Environmental Protection Agency.
NORAZIAN, M. N., SHUKRI, A., YAHAYA, P. M., AZAM, N., FITRI, N. F. & YUSOF, M. 2013. Roles of imputation methods for filling the missing values: A review.
NORAZIAN, M. N., SHUKRI, Y. A. & AZAM, R. N. 2008. Estimation of missing values in air pollution data using single imputation techniques.
OBERDÖRSTER, G., OBERDÖRSTER, E. & OBERDÖRSTER, J. 2005. Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. *Environmental health perspectives***,** 823-839.
OFFICE OF ENVIRONMENT & HERITAGE 2012. National Environment Protection (Ambient Air Quality) Measure: New South Wales Annual Compliance Report 2012. Office of Environment & Heritage.

OFFICE OF ENVIRONMENT & HERITAGE. 2015. *Air quality monitoring network quality assurance* [Online]. Available: http://www.environment.nsw.gov.au/aqms/qualityassurance.htm [Accessed 11th May 2017].

PARK, K., CHOW, J. C., WATSON, J. G., TRIMBLE, D. L., DORAISWAMY, P., PARK, K., ARNOTT, W. P., STROUD, K. R., BOWERS, K. & BODE, R. 2006. Comparison of continuous and filter-based carbon measurements at the Fresno Supersite. *Journal of the Air & Waste Management Association,* 56**,** 474-491.

PRICE, M., BULPITT, S. & MEYER, M. B. 2003. A comparison of PM 10 monitors at a Kerbside site in the northeast of England. *Atmospheric Environment,* 37**,** 4425-4434.

R DEVELOPMENT CORE TEAM 2011. *R: A Language and Environment for Statistical Computing,* Vienna, Austria, The R foundation for Statitical Computing.

RIZZO, M., SCHEFF, P. A. & KALDY, W. 2003. Adjusting tapered element oscillating microbalance data for comparison with Federal Reference Method PM2. 5 measurements in Region 5. *Journal of the Air & Waste Management Association,* 53**,** 596-607.

ROSAMOND, M. S., THUSS, S. J. & SCHIFF, S. L. 2012. Dependence of riverine nitrous oxide emissions on dissolved oxygen levels. *Nature geoscience,* 5**,** 715-718.

RUPPRECHT & PATASCHNICK, C. I. 1993. Technical Bulletin to End-Users of TEOM Series 1400 PM-10 Monitors re: Low Temperature Operation.

RUPPRECHT & PATASCHNICK, C. I. 2008. TEOM Series 1400a Ambient Particulate (PM-10) Monitor: Operating Manual.

SALVADOR, C. M. & CHOU, C. C.-K. 2014. Analysis of semi-volatile materials (SVM) in fine particulate matter. *Atmospheric Environment,* 95**,** 288-295.

SCHWAB, J. J., FELTON, H. D., RATTIGAN, O. V. & DEMERJIAN, K. L. 2006. New York state urban and rural measurements of continuous PM2. 5 mass by FDMS, TEOM, and BAM. *Journal of the Air & Waste Management Association,* 56**,** 372-383.

SCHWEIZER, D., CISNEROS, R. & SHAW, G. 2016. A comparative analysis of temporary and permanent beta attenuation monitors: The importance of understanding data and equipment limitations when creating PM 2.5 air quality health advisories. *Atmospheric Pollution Research,* 7**,** 865-875.

SIENFELD, J. H. & PANDIS, S. N. 1998. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change,* New York, John Wiley & Sons, Inc.

SLOANE, C. S., WATSON, J., CHOW, J., PRITCHETT, L. & RICHARDS, L. W. 1991. Size-segregated fine particle measurements by chemical species and their impact on visibility impairment in Denver. *Atmospheric Environment. Part A. General Topics,* 25**,** 1013-1024.

SOFUOGLU, S. C., SOFUOGLU, A., BIRGILI, S. & TAYFUR, G. 2006. Forecasting ambient air SO2 concentrations using artificial neural networks. *Energy Sources, Part B,* 1**,** 127-136.

TANG, H., LEWIS, E., EATOUGH, D., BURTON, R. & FARBER, R. 1994. Determination of the particle size distribution and chemical composition of semi-volatile organic compounds in atmospheric fine particles with a diffusion denuder sampling system. *Atmospheric Environment,* 28**,** 939-947.

THERMO SCIENTIFIC. 2014. Model 5014i Beta: Instruction Manual.

THOMAS, S. & JACKO, R. B. 2007. Model for forecasting expressway fine particulate matter and carbon monoxide concentration: application of regression and neural network models. *Journal of the Air & Waste Management Association,* 57**,** 480-488.

TSIGARIS, P., SCHEMENAUER, R 2014. Reconstructing the Historic Database of Annual PM2.5 Values for Kamloops, B.C. by
Calculating the Offset between TEOM and BAM Measurements.
UNITED STATES ENVIRONMENTAL PROTECTION AGENCY. 1997. *Guidance for network design and optimum site exposire for PM2.5 and PM10* [Online]. Available: http:// www.epa.gov/ttn/amtic/files/cfr/recent/pmnaaqs.pdf [Accessed 20/09/2017].
WARK, K., WARNER, C. F. & DAVIS, W. T. 1998. *Air Pollution: Its Origin and Control,* Menlo Park, California, Addison Wesley Longman.
WATSON, J. G. & CHOW, J. C. 2002. Comparison and evaluation of in situ and filter carbon measurements at the Fresno Supersite. *Journal of Geophysical Research: Atmospheres,* 107.
WILLMOTT, C. J., ROBESON, S. M. & MATSUURA, K. 2012. A refined index of model performance. *International Journal of Climatology,* 32**,** 2088-2094.
WORLD HEALTH ORGANIZATION 2013. IARC: Outdoor air pollution a leading environmental cause of cancer deaths. *In:* CANCER, I. A. F. R. O. (ed.).
WORLD HEALTH ORGANIZATION 2016. Ambient Air Pollution: A global assessment of exposire and burden of disease. Geneva.

Appendix 1: Overview of the main characteristics of the continuous TEOM and BAM samplers.

Table AP1- 1. Overview of the main characteristics of the two continuous samplers.

Sampler	Model	Operated by	In use since	Sample flow rate ($L.min^{-1}$)	Sample flow heating	Pre-separation system	Measuring principle	Measuring range ($ug.m^{-3}$)
TEOM	R&P Thermo Fisher TEOM 1400AB	Office of Environment and Heritage.	17/01/2003	16.7[a] 3.1[b]	50 degrees Celcius	Impactor (US-EPA 40 CFR 50, App. L)	Micro-balance	5 to >10,000
BAM	Met-One BAM 5014i	Office of Environment and Heritage.	29/11/2012	16.7	Ambient +- 5 degrees Celcius	VSCC	Beta-ray attenuation	4-10,000

[a] Total sample flow pulled though the impactor.
[b] Sample flow directed to measuring chamber after passing the flow splitter.

Appendix 2: Summary of literature used to correct TEOM measurements.

Table AP2- 1. Summary of literature of methods used to adjust the TEOM.

Author	Title	Instrument	Location	PM Measured	Key Findings	Proposed correction method	Comments
Correction using chemical speciation data.							
Chung et al. 2001.	Comparison of Real-Time Instruments Used To Monitor Airborne Particulate Matter	TEOM, BAM, FRM.	Bakersfield, California.	$PM_{2.5}$, PM_{10}.	• $PM_{2.5}$ TEOM readings were lower than collocated $PM_{2.5}$ FRM. There is a statistically significant relationship between this difference and NO_3- concentrations. • BAM was not heavily influenced by meteorological conditions and particle composition. • Use of the correction method significantly improved agreement between the TEOM and FRM.	The amount of ammonium nitrate (N_4NO_3) present in $PM_{2.5}$ was converted to $\mu g/m^3$ and added to the raw TEOM measurements. This provided an improved agreement between the TEOM and the FRM, insinuating that the error observed in the TEOM is in line with the NH_4NO_3 concentration. The remaining error may be due to the loss of other volatiles being evaporated, such as organic compounds.	Do not have particle ammonium nitrate particle composition data.
Godri et al. 2009.	Evaluation and application of a semi-continuous chemical characterization system for water soluble inorganic $PM_{2.5}$ and associated precursor gases	TEOM, Dichot, Partisol.	Toronto, Canada.	$PM_{2.5}$.	• Difference exists between TEOM and Dichot filter measured $PM_{2.5}$, chiefly during winter. This difference is attributed to the volatization of nitrate in the TEOMs heated inlet air stream. • The TEOM was calibrated to the Dichot filter using nitrate data to normalize the variation.	Due to loss of volatile material from the TEOM, the TEOM was calibrated to the NAPS dichot filter.	Need chemical composition data to apply this correction method.
Hauck et al. 2004.	On the equivalence of gravimetric PM data with TEOM and	TEOM, BAM, gravimetric methods.	Austria.	$PM_{2.5}$.	• Fair agreement of TEOM, BAM and high volume sampling when grouped by temperature and chemical composition.	Assuming that all nitrate has been volatilized from the heated inlet for the TEOM, the nitrate measured on the filters are added to the TEOM PM concentration	Do not have nitrate particle composition data.

	beat-attenuation measurements.				• The TEOMs and BAMs are not significantly different from each other in the summer period. Winter months show large discrepancies in $PM_{2.5}$. • Low nitrate, occurring mostly in summertime, is correlated with high TEOM values. • When there is a low nitrate content on the gravimetric filters, there is a good agreement between the gravimetric and the continuous monitors. • Correcting for nitrate loss improved the agreement for winter data and for higher nitrate concentrations.	data as ammonium nitrate. Regression line and R^2 value improve considerably.	
Lee et al. 2005.	Evaluation of continuous and filter based methods for measuring $PM_{2.5}$ mass concentration.	TEOM, CAMM, RAMS.	Houston and Seattle, America.	$PM_{2.5}$	• Difference existed at some sites between the RAMS and the TEOM. This difference probably a result of ammonium nitrate loss and water vapor. • Difference from 24hour averaged continuous $PM_{2.5}$ and 24hour integrated $PM_{2.5}$ is likely due to the loss of semi-volatile material.	Difference in PM concnetrations may be due to the loss of ammonium nitrate. Hence, add lost ammonium nitrate to $PM_{2.5}$ readings and examine change in agreement.	Don't have access to particle composition data.
Li et al. 2012.	Field evaluation of particulate matter measurements using tapered element oscillating microbalance in a layer house.	TEOM, FRM.		PM_{10}, $PM_{2.5}$.	• The TEOM read lower PM_{10} and $PM_{2.5}$ readings than the gravimetric method. • Significantly higher PM mass concentrations were measured at lower internal temperature settings of the instrument. • Regression analyses used to estimate the effects of the predictor variable on the response. • Adding NH_4NO_3 to the TEOM $PM_{2.5}$ concentrations did not significantly improve the relationship between the TEOM and filter based methods. Therefore, NH_4NO_3 was insignificant	Adding NH_4NO_3 to the TEOM $PM_{2.5}$ did not significantly improve the relationship between TEOM and filter-based methods. Making TEOM and FRM measurements comparable remains a big challenge.	Do not have any NH_4NO_3 data.

					in its contribution to the PM mass. Hence, a substantial portion of mass loss may have been from the volatization of PM bound moisture and VOCs/SVOCs.		
Schwab et al. 2006.	New York State urban and rural measurements of continuous $PM_{2.5}$ mass by FDMS, TEOM and BAM.	FDMS TEOM, TEOM, BAM, FRM.	New York.	$PM_{2.5}$.	• BAM and FDMS TEOM are highly correlated, and are ~25% higher than the FRM filter measurements at one site. • Mass reconstruction of the network filter data is completed to examine the contribution of volatile species to the $PM_{2.5}$ mass.	Use speciation methods to reconstruct $PM_{2.5}$.	Do not have access to chemical speciation data.
Zhu, Zhang and Lioy, 2007.	Evaluation and comparison of continuous fine particulate matter monitors for measurement of ambient aerosols	FRM, Nephelometers, TEOM and BAM.	New Jersey, America.	$PM_{2.5}$.	• Two TEOMs (TEOM 1400 and TEOM FDMS) correlated well (r^2 0.85), and the two BAMs exhibiting a weaker correlation (r^2 0.6). Seasonal differences expressed in the TEOM, which is a result of the semi-volatile material loss in the winter.	Study was just a comparison between devices. But results suggest TEOM measurements needs to account for semi-volatile material loss.	Don't have access to particle composition data.
Correction using correction factors							
Green and Barratt. 2001.	Evaluation of TEOM correction factors for assessing the EU Stage 1 limit values for PM_{10}.	TEOM, Partisol.	London.	PM_{10}.	• The degree of seasonal variability may change on a yearly basis depending on meteorological conditions. Therefore, any correction factors applied to TEOM data should integrate the local geographical and temporal variability .	Single correction factors applied to correct data don't produce accurate results. Correction factors calculated seasonally or annually results in better agreement than a single correction factor applied over the whole period.	It is best to calculate a correction method that accounts for local geography and temporal variability.
Tsigaris & Schemenauer. 2014.	Reconstructing the Historic Database of Annual $PM_{2.5}$ Values for Kamloops. B.C. by	TEOM, BAM.	Kamloops, B.C.	$PM_{2.5}$.	• The annual average $PM_{2.5}$ TEOM (data from 1998 to 2010) were underestimated. • Monthly BAM average is always higher than the corresponding monthly TEOM average. • Simple mean adjustment method used.	Applying an annual mean adjustment method, with the average adjustment factor ranging from 3.1 ug/m^3 to 3.3ug/m^3. This adjustment factor is added to the TEOM values to create a $PM_{2.5}$ series that can be merged with the	This adjustment method may be appropriate for annually averaged data. However, given our data is recorded in hourly intervals, and we know that there are factors such

	Calculating the Offset between TEOM and BAM Measurements.				• The results from the adjustment indicate that the city has commonly exceeded annual mean values of $PM_{2.5}$ that are above the provincial guideline of 8 $\mu g/m^3$ since 1998.	modern BAM instrument measurements.	as temperature and relative humidity that influence the readings depending on the instrument, we should first investigate using these suite of variables available to correct the TEOM, before we look at applying a simple approach to annual averages.
Winkel et al, 2015.	Equivalence testing of filter based, beta-attenuation, TEOM and light-scattering devices for measurement of PM_{10} concentration in animal houses.	TEOM, reference sampler.	Wagenungen, the Netherlands.	PM_{10}.	• TEOM underestimated the European Reference Sampler (RES) concentration at all four sampling sites. The mean underestimation varied from 21% (at the office space) to 33% (in pigs).	Duplicate sampling can be employed to reduce random errors related to differences between samplers, whereas correction factors (specific to the level of animal categories or animal housing systems) can be determined and applied to reduce systematic deviations from a reference sampler.	Single correction factors not appropriate generally (Green et al. 2001). However, it was suitable in this case possibly for two reasons; the composition of PM in the animal house may be homogeneous and low in ammonium nitrate when compared to ambient PM. Secondly, pig houses are kept insulated, containing ventilation systems that maintain the temperature and relative humidity within certain limits.
Wu, J, Miner, AM & Delfino, 2006.	Exposure assessment of particulate matter air pollution before, during and after the 2003 Southern Californian wildfires.	BAM, TEOM and gravimetric methods.	Southern California, America.	$PM_{2.5}$, PM_{10}.	• BAM instrument over-estimated $PM_{2.5}$ concentrations compared to the filter based methods. • Significant differences present between filter based vs TEOM and BAM instruments. • Using PM data from different samplers may cause issues when estimating PM.	$PM_{2.5}$ measurements should be transformed to a single standard. However, these are site specific and based on the particle composition. Sometimes appropriate conversions cannot be calculated for particular sites and instruments. Gravimetric data, real-time data, and satellite data can be used to predict PM concentrations.	Investigate developing correction method for this site. Do not have time to investigate incorporating satellite data into predictions. This would be suitable for future research.

Modelling							
Bilonick et al. 2015.	Using structural equation modeling to construct calibration equations relating $PM_{2.5}$ mass concentration samplers to the federal reference method sampler.	FRM, TEOM and speciation samplers.	Pittsburgh, Pennsylvania, America.	$PM_{2.5}$.	• TEOM imprecision and TEOM bias (relative to the FRM) decreased as temperature increased. • Calibration model developed to link the TEOM to the FRM and speciation devices, as a function of temperature. • Modelling demonstrated that the FRM samplers were more precise than the TEOM and speciation devices, and the TEOM displayed negative bias towards the FRM. • Ordinary regression assumes the response variable is free from random error, though this is not always the case. Although, ordinary regression provides a calibration that is nearly correct, when one instrument is a lot more precise than the other.	Structural equation modelling was utilized to relate the TEOM to the FRM and speciation samplers as a function of ambient temperature.	Possible to be ultilised in our research.
Gehrig et al. 2005.	A New Method to Link PM_{10} concentrations from automatic monitors to the manual gravimetric reference method according to EN12341.	BAM, FRM.	Switzerland.	PM_{10}.	• Linear regression results indicate good agreement of the means of the corrected data set. • Day-to-day correction was applied, and produced excellent agreement of annual means, great correlation and a reduction in the standard deviation of differences.	Day-to-day correction used in the study include: *Equation 1:* calculates daily correction factors corresponding to the ratio gravimetry/monitor for those days which a gravimetric value was measured. *Equation 2:* for days without gravimetric measurements corresponding to the mean of the ratios gravimetry/monitor of the two nearest days with gravimetry data.	Answer may lie in investigating ratios. Look into this when developing methods to correct TEOM. Develop a procedure that can account for changes in meteorological conditions, or of the aerosols composition, instead of relying on long term comparisons.
Hsu et al. 2016.	Collocated comparisons of continuous and filter-based $PM_{2.5}$ measurements	TEOM, SHARP (FEM), Partisol.	Alberta, Canada.	$PM_{2.5}$.	• Hourly TEOM $PM_{2.5}$ were consistently ~20-50% lower than that of SHARP. • Orthogonal regression equations were derived from FRM and TEOM to adjust the TEOM values, and improve	Orthogonal regression equations to correct historical TEOM data, to examine long term trends within the network.	Could apply orthogonal regression for model.

	at Fort McMurray, Alberta, Canada.				its agreement with FRM, especially for the cold season. • These adjusted measurements enable a long term trend analysis of the network.		
Kashuba and Scheff, 2008.	Non-linear regression adjustments of multiple continuous monitoring methods produce effective characterization of short-term fine particulate matter.	TEOM, BAM, Nephelometer, FRM.	United States.	$PM_{2.5}$.	• Least squares regression and non-linear regression using meteorological variables are used in model. • Nonlinear models have higher correlation than linear models when used on the same data. But the variation in correlation is not always going to be significantly better. So there is tradeoffs between simplicity of a model and degree of statistical association.	Apply linear regression model or non-linear regression model.	Both linear and non-linear models are worth considering.

Appendix 3: Graphical data to test assumptions of ARDL model.

Transforming for linearity

X-variables whose linearity did not improve when plotted with the transformed BAM are shown in Figure AP3-1, and include Ozone and SO_2. *X*-variables that were transformed to improve linearity between them and transformed BAM are shown in Figure AP3-2, and include NEPH, CO, NO_x, NO and NO_2.

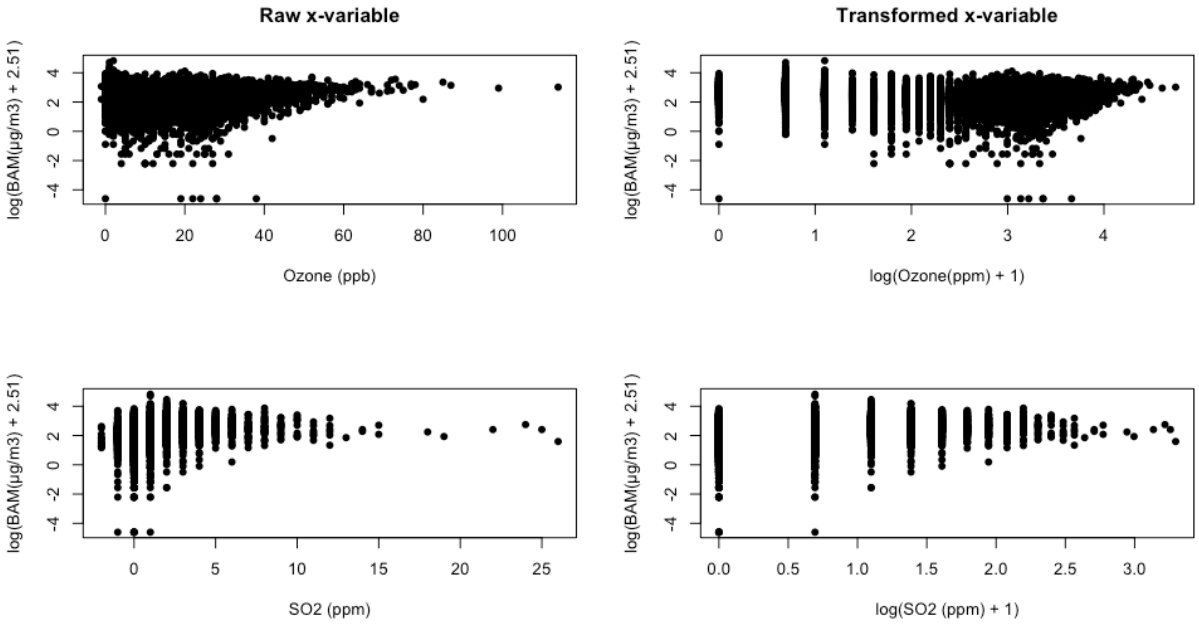


Figure AP3- 1. Plots on the left hand side shown transformed BAM against untransformed *x*-variables. Plots on the right hand side show transformed BAM against transformed *x*-variables. The linearity of the relationship does not improve when transformed in these cases.

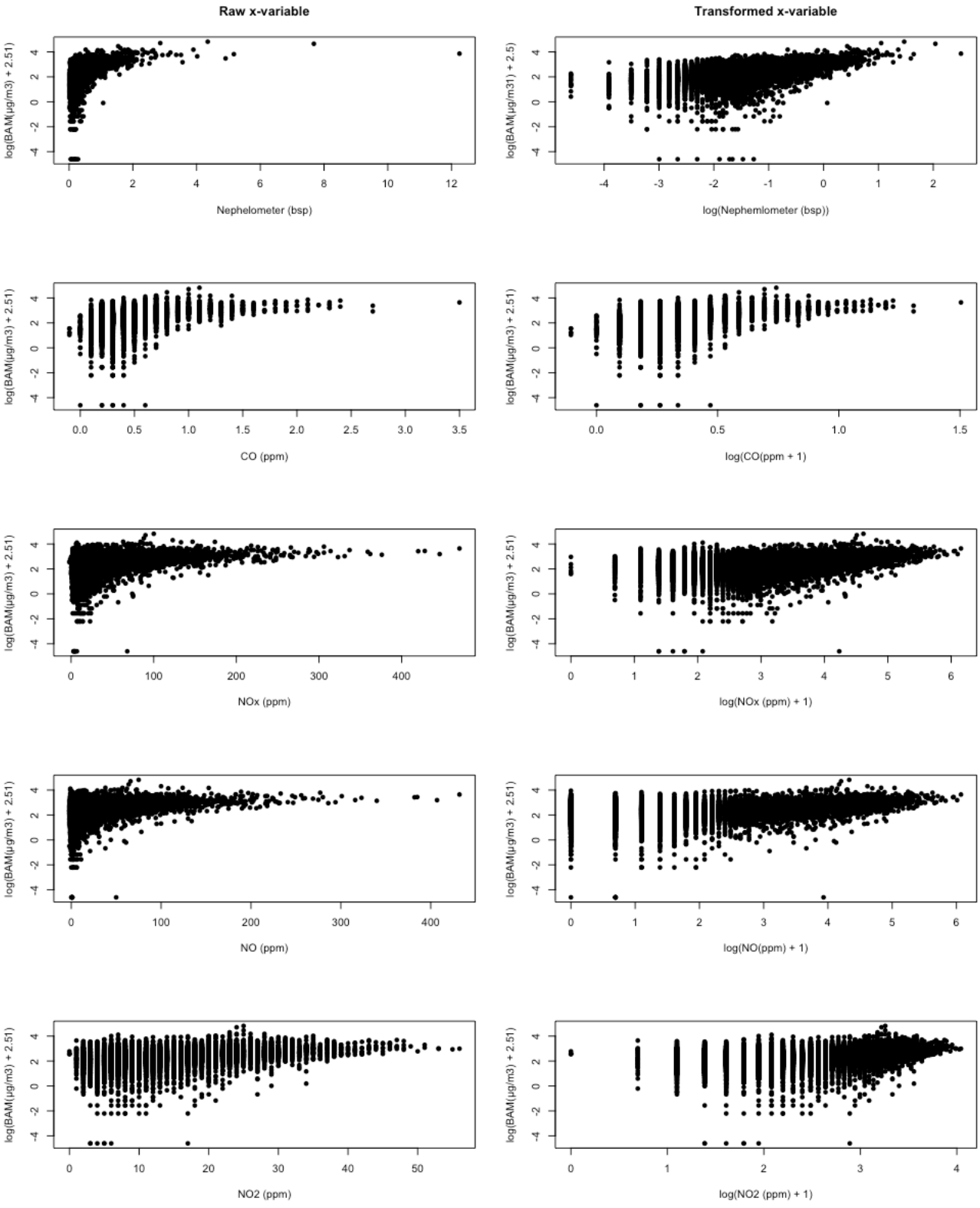


Figure AP3- 2. Plots on the left hand side show transformed BAM against untransformed x-variables. Plots on the right hand column show transformed BAM against transformed x-variables. The linearity of the relationship does improve in these cases.

Testing for multicollinearity

Before carrying out a thorough investigation of instrument error, it is necessary to determine possible multicollinearity between independent variables. Table AP3- 1 summarises the correlation coefficients between hourly values. The results indicate that a high degree of multicollinearity exists between the concentrations of NO_x and NO_2. This dependency exists due to the fact that NO_2 is a part of the make-up NO_x. Additionally, NO_x and CO have a high correlation coefficient of 0.79. Therefore, only NO_x or CO should be included in the model. Furthermore, NEPH, NEHP lag 1 and NEPH lag 2 all have a high correlation coefficient, ≥ 0.8. Therefore, only one of these variables can be included in the final model.

Table AP3- 1. Cross correlation matrix showing the correlation between variables, for hourly data. Variables with a correlation of rho ≥ 0.6 are highlighted in yellow, indicating that caution should be used if using both of these parameters in a model as they may possess multicollinearity. Variables with a correlation of rho ≥ 0.8 are highlighted in red. These pairs should not be used in a model together as they will definitely produce overfitting as a result of multicollinearity. The rho of BAM, TEOM, NEPH, PM_{10} and gasses are calculated from transformed values.

	BAM	**TEMP**	**RH**	**TEOM**	**TEOM lag 1**	**TEOM lag 2**	**TEOM lag 24**	**NEPH**	**NEPH lag 1**	**NEPH lag 2**	**NEPH lag 24**	**PM_{10}**	**PM_{10} lag 1**	**PM_{10} lag 2**	**PM_{10} lag 24**	**CO**	**NO**	**NOx**	**NO2**	**Wind Speed**
BAM	1.00	-0.09	0.17	0.49	0.54	0.54	0.29	0.56	0.59	0.57	0.31	0.39	0.42	0.43	0.19	0.41	0.30	0.35	0.32	-0.18
TEMP	-0.09	1.00	-0.40	0.00	0.00	0.03	0.01	-0.12	-0.10	-0.06	-0.10	0.10	0.11	0.14	0.08	-0.21	-0.41	-0.45	-0.42	0.28
RH	0.17	-0.40	1.00	0.20	0.20	0.16	0.14	0.42	0.38	0.31	0.25	-0.07	-0.08	-0.11	-0.04	0.45	0.35	0.41	0.37	-0.46
TEOM	0.49	0.00	0.20	1.00	0.76	0.61	0.38	0.76	0.64	0.54	0.30	0.74	0.59	0.47	0.27	0.60	0.45	0.52	0.47	-0.30
TEOM lag 1	0.54	0.00	0.20	0.76	1.00	0.76	0.37	0.72	0.76	0.64	0.31	0.58	0.74	0.59	0.26	0.55	0.41	0.46	0.42	-0.26
TEOM lag 2	0.54	0.03	0.16	0.61	0.76	1.00	0.32	0.64	0.72	0.76	0.30	0.47	0.58	0.74	0.23	0.46	0.33	0.37	0.34	-0.22
TEOM lag 24	0.29	0.01	0.14	0.38	0.37	0.32	1.00	0.34	0.32	0.29	0.76	0.26	0.24	0.21	0.74	0.28	0.17	0.21	0.21	-0.13
NEPH	0.56	-0.12	0.42	0.76	0.72	0.64	0.34	1.00	0.91	0.81	0.43	0.60	0.55	0.48	0.24	0.65	0.49	0.56	0.51	-0.41
NEPH lag 1	0.59	-0.10	0.38	0.64	0.76	0.72	0.32	0.91	1.00	0.91	0.43	0.51	0.60	0.55	0.23	0.58	0.45	0.49	0.44	-0.36
NEPH lag 2	0.57	-0.06	0.31	0.54	0.64	0.76	0.29	0.81	0.91	1.00	0.42	0.43	0.51	0.60	0.22	0.49	0.39	0.41	0.36	-0.31
NEPH lag 24	0.31	-0.10	0.25	0.30	0.31	0.30	0.76	0.43	0.43	0.42	1.00	0.20	0.20	0.18	0.60	0.31	0.20	0.25	0.24	-0.20
PM0	0.39	0.10	-0.07	0.74	0.58	0.47	0.26	0.60	0.51	0.43	0.20	1.00	0.76	0.62	0.32	0.43	0.41	0.41	0.31	-0.06
PM_{10}lag 1	0.42	0.11	-0.08	0.59	0.74	0.58	0.24	0.55	0.60	0.51	0.20	0.76	1.00	0.76	0.31	0.38	0.34	0.34	0.26	-0.03
PM_{10} lag 2	0.43	0.14	-0.11	0.47	0.59	0.74	0.21	0.48	0.55	0.60	0.18	0.62	0.76	1.00	0.28	0.29	0.25	0.25	0.20	0.00
PM_{10} lag 24	0.19	0.08	-0.04	0.27	0.26	0.23	0.74	0.24	0.23	0.22	0.60	0.32	0.31	0.28	1.00	0.16	0.13	0.14	0.12	0.00
CO	0.41	-0.21	0.45	0.60	0.55	0.46	0.28	0.65	0.58	0.49	0.31	0.43	0.38	0.29	0.16	1.00	0.74	0.79	0.65	-0.49
NO	0.30	-0.41	0.35	0.45	0.41	0.33	0.17	0.49	0.45	0.39	0.20	0.41	0.34	0.25	0.13	0.74	1.00	0.91	0.69	-0.46
NOx	0.35	-0.45	0.41	0.52	0.46	0.37	0.21	0.56	0.49	0.41	0.25	0.41	0.34	0.25	0.14	0.79	0.91	1.00	0.91	-0.58
NO2	0.32	-0.42	0.37	0.47	0.42	0.34	0.21	0.51	0.44	0.36	0.24	0.31	0.26	0.20	0.12	0.65	0.69	0.91	1.00	-0.60
Wind Speed	-0.18	0.28	-0.46	-0.30	-0.26	-0.22	-0.13	-0.41	-0.36	-0.31	-0.20	-0.06	-0.03	0.00	0.00	-0.49	-0.46	-0.58	-0.60	1.00

Appendix 4: Blocking month and hour input variables.

Data blocking was carried out on monthly and hourly data to prevent an excess number of predictor variables in the model. A model was created using all of the input variables, with the summary of the model printed below. Cut-offs for each block was decided based on the significance level, through the *p-value*, and the estimate and *t value*.

For the hourly data, the first block, *a*, consists of hours 11:00 p.m., 12:00 a.m., 1:00 a.m. and 2:00 a.m. The starting point, 11:00 p.m. is strongly significant, and remains that way until 1:00 a.m. 2:00 a.m. is less significant, with a *p-value* of 0.09. This was the cut-off for the first block as we know that the TEOM and BAM do behave very differently from around midnight to 7:00 a.m., so we wanted to try and explain this as best as possible, having two blocks over this time. The next block, *b*, is from 3:00 a.m. to 7:00 a.m. The cut-off for block *b* was 7:00 a.m. as the 7th hour is not significant. We chose to have the 7th hour in this block, rather than the next one, to keep only positive estimates and *t-values* in this block. The next block, *c*, runs from 8:00 a.m. to 3:00 p.m. In this block, there are some significant and not significant hours present. We consciously chose to keep this block as a whole, and not broken down into smaller blocks, for the sake of ensuring a parsimonious model was developed. There were only two strongly significant hours over this time period, 9:00 a.m. and 3:00 p.m., as indicated by the three asterisks beside the *p-values*. Lastly, the fourth block, *d*, runs from 4:00 p.m. till 10:00 p.m., containing 3 strongly significant hours.

The monthly data too needs to be broken into blocks. Upon examining the output below, we decided to block the hourly data into two blocks. The first block, *a*, runs from November to March. The second block, *b*, runs from April to October. These cut-off were chosen based on the *p-values*, and they also fit well with keeping positive and negative estimate and *t-values* in separate blocks.

```
Call:
lm(formula = baml ~ rh + temp + hr + mth + teoml + teoml.l1 +
    teoml.l2 + teoml.l24 + nephl.l1 + pm10l.l1 + pm10l.l2 + pm10l.l2 +
    pm10l.l24 + lco + lno + lno2 + wd + ws, data = mds)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8021 -0.1461  0.0512  0.2272  1.8394

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.624e+00  1.011e-01  25.964  < 2e-16 ***
rh          -6.067e-03  3.648e-04 -16.630  < 2e-16 ***
temp        -1.631e-02  1.571e-03 -10.380  < 2e-16 ***
hr1          8.703e-02  2.585e-02   3.367 0.000761 ***
hr2          3.458e-01  2.101e-01   1.646 0.099859 .
hr3          1.941e-01  2.638e-02   7.359 1.95e-13 ***
hr4          1.796e-01  2.579e-02   6.966 3.40e-12 ***
hr5          1.601e-01  2.575e-02   6.218 5.16e-10 ***
hr6          1.204e-01  2.582e-02   4.662 3.16e-06 ***
hr7          2.358e-02  2.592e-02   0.910 0.362853
hr8         -6.857e-02  2.585e-02  -2.653 0.007995 **
hr9         -1.243e-01  2.622e-02  -4.739 2.16e-06 ***
hr10        -6.050e-02  2.692e-02  -2.247 0.024638 *
hr11        -7.014e-03  2.760e-02  -0.254 0.799422
hr12         4.685e-03  2.799e-02   0.167 0.867074
hr13        -6.689e-02  2.834e-02  -2.360 0.018281 *
hr14        -6.101e-02  2.837e-02  -2.150 0.031536 *
hr15        -1.003e-01  2.814e-02  -3.564 0.000366 ***
hr16        -5.277e-02  2.765e-02  -1.908 0.056371 .
hr17         4.156e-02  2.705e-02   1.536 0.124496
hr18         1.945e-01  2.658e-02   7.315 2.69e-13 ***
hr19         1.933e-01  2.636e-02   7.334 2.34e-13 ***
hr20         1.271e-01  2.611e-02   4.869 1.13e-06 ***
hr21         4.904e-02  2.576e-02   1.904 0.056992 .
hr22         3.789e-02  2.550e-02   1.486 0.137397
hr23         1.286e-02  2.543e-02   0.506 0.613003
mth2         3.364e-02  1.991e-02   1.690 0.091139 .
mth3         7.267e-02  2.026e-02   3.588 0.000335 ***
mth4        -2.645e-02  2.320e-02  -1.140 0.254263
mth5        -6.427e-02  2.525e-02  -2.545 0.010922 *
mth6        -8.509e-02  2.624e-02  -3.243 0.001187 **
mth7        -9.472e-02  2.746e-02  -3.450 0.000563 ***
mth8        -8.514e-02  2.639e-02  -3.226 0.001256 **
mth9        -7.871e-02  2.392e-02  -3.290 0.001002 **
mth10       -6.077e-02  2.133e-02  -2.849 0.004385 **
mth11        1.442e-02  1.951e-02   0.739 0.459863
mth12        9.921e-02  2.021e-02   4.910 9.21e-07 ***
teoml        9.106e-02  1.156e-02   7.878 3.53e-15 ***
teoml.l1     5.277e-02  1.669e-02   3.162 0.001572 **
teoml.l2     1.327e-01  1.477e-02   8.990  < 2e-16 ***
teoml.l24    7.745e-02  8.962e-03   8.641  < 2e-16 ***
nephl.l1     3.657e-01  1.099e-02  33.284  < 2e-16 ***
pm10l.l1    -1.483e-01  2.458e-02  -6.036 1.62e-09 ***
pm10l.l2     1.458e-01  2.349e-02   6.206 5.57e-10 ***
pm10l.l24   -2.183e-02  1.319e-02  -1.654 0.098068 .
lco          4.886e-01  5.398e-02   9.050  < 2e-16 ***
lno         -1.010e-02  5.805e-03  -1.741 0.081785 .
lno2         4.247e-02  1.208e-02   3.517 0.000438 ***
wd          -1.078e-04  4.704e-05  -2.292 0.021916 *
ws           4.841e-02  4.136e-03  11.706  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4673 on 15765 degrees of freedom
  (3836 observations deleted due to missingness)
Multiple R-squared:  0.4417,  Adjusted R-squared:  0.4399
F-statistic: 254.5 on 49 and 15765 DF,  p-value: < 2.2e-16
```

Appendix 5: Measures used to determine the most appropriate model.

Adjusted R^2, $\bar{R}^2$.

Regular R^2 is not suitable to determine the predictive ability of a model, as adding any variable typically increases the R^2 value. If one were to keep adding variables until the highest R^2 value is achieved, we would likely be left with a model that is not parsimonious. A way to overcome this issue is to examine the adjusted R^2 value, and can be calculated using the following formula:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{N\ -\ 1}{N\ -\ k\ -\ 1},$$

where *N* is the number of observations and *k* is the number of predictors. This is more suitable then the R^2 as it does not increase when an independent variable is added. The model with the biggest $\bar{R}^2$ is the best model. Increasing the $\bar{R}^2$ results in a decreasing estimate of variance of the forecast errors:

$$\hat{\sigma}^2 = \frac{SSE}{N - k - 1},$$

where $\hat{\sigma}^2$ is the estimate of variance, SSE is the sum of squared errors.

Cross-validation, *CV.*

Cross-validation (CV) is an effective way to assess the predictability of a model. The following procedure is used to assess the predictive accuracy of the model:

a) Remove observation *i* from the dataset, and build the model on the remaining observations. Compute the error ($e_i * = \ y_i - \ \hat{y}_i$) for the observation that was omitted.
b) Repeat step a) for *i=1,...,n.*
c) Calculate the MSE from $\hat{e}_i *$, …, $\hat{e}_n *$. This is called the CV.

The model with the smallest CV is the best.

Akaike's Information Criterion, *AIC.*

An alternative method to examine the predictive ability of a model is Akaike's Information Criterion (AIC). The AIC is calculated as follows:

$$AIC = N \log\left(\frac{SSE}{N}\right) + 2(k + 2),$$

The model possessing the lowest AIC is the best model.

Schwarz Bayesian Information Criterion, *BIC.*

Lastly, Schwarz Bayesian Information Criterion (BIC) can also be used to assess the predictive ability of a model. The BIC is calculated as follows:

$$BIC = N\ log\left(\frac{SSE}{N}\right) + (k+2)log\ (N).$$

Minimizing the BIC provides the best model. The model selected by the BIC is either the same as the AIC or one with fewer terms included, as the BIC penalizes according to the number of parameters in the model.

Appendix 6: Specifics of methods used for variable selection for predictive model.

Stepwise regression

Given that there a large number of variables, it is not suitable to construct all possible models and check the measures of predictive accuracy on each of these models. Another technique is needed to limit the number of models investigated. We employed both manual f-test forward and backward selection to assist with variable selection.

Manual f-test backward selection

This method works by making a model containing all of the potential predictors, namely, *lmfull*. Next, the `drop1` function is used, with the test set to equal 'f', producing an output showing the degrees of freedom, sum of squares, RSS, AIC, F value and P value. From this output, insignificant variables are identified and removed, then the code is re-run, and the output is re-examined for insignificant variables. The code used is shown below:

```
#set linear model containing all possible variables available for
selection
lmfull <- lm(baml ~
temp+rh+hrbk+mthbk+teoml+teoml.l1+teoml.l2+teoml.l24+nephl.l1+nephl.l24+pm
10l+pm10l.l1+pm10l.l2+pm10l.l24+lco+lno2+lno+ws+wdir, data=mds)
#use the drop1 function to show model output
drop1(lmfull, test = "F")
#nephl.l24 was identified as the least significant, so minus this from the
model
drop1(update(lmfull, ~ . -nephl.l24), test = "F")
```

Given that we are after a parsimonious model, NEPH lag 24 was removed from the model as it was not significant (p=0.06). The model was re-run, and the output was examined. This time, PM_{10} was removed as it too was no longer slightly significant (p=0.07). Then wind direction and PM10 lag 24 were removed for their low level of significance (p=0.02 for both variables). After this, all variables were strongly significant at a significance level of 0.05. The final model using the f-test backward selection is shown below.

```
Call:
lm(formula = baml ~ temp + rh + hrbk + mthbk + teoml + teoml.l1 +
    teoml.l2 + teoml.l24 + nephl.l1 + pm10l.l1 + pm10l.l2 + lco +
    lno2 + lno + ws, data = mds)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9576 -0.1576  0.0480  0.2330  1.8599

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.5424929  0.0885141  28.724  < 2e-16 ***
temp        -0.0146844  0.0011929 -12.310  < 2e-16 ***
rh          -0.0054805  0.0003288 -16.667  < 2e-16 ***
hrbkb        0.1060767  0.0140436   7.553 4.47e-14 ***
hrbkc       -0.0841233  0.0142143  -5.918 3.32e-09 ***
hrbkd        0.0593728  0.0135268   4.389 1.14e-05 ***
mthbkb      -0.1041434  0.0118872  -8.761  < 2e-16 ***
teoml        0.0864400  0.0113426   7.621 2.66e-14 ***
teoml.l1     0.0537523  0.0165574   3.246  0.00117 **
teoml.l2     0.1343761  0.0144373   9.308  < 2e-16 ***
teoml.l24    0.0600673  0.0071617   8.387  < 2e-16 ***
nephl.l1     0.3592783  0.0105505  34.053  < 2e-16 ***
pm10l.l1    -0.1537244  0.0244864  -6.278 3.52e-10 ***
pm10l.l2     0.1518855  0.0233496   6.505 8.01e-11 ***
lco          0.5091722  0.0522786   9.740  < 2e-16 ***
lno2         0.0539037  0.0112575   4.788 1.70e-06 ***
lno         -0.0213109  0.0054326  -3.923 8.79e-05 ***
ws           0.0443205  0.0040665  10.899  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4714 on 15836 degrees of freedom
  (3797 observations deleted due to missingness)
Multiple R-squared:  0.4306,  Adjusted R-squared:   0.43
F-statistic: 704.5 on 17 and 15836 DF,  p-value: < 2.2e-16
```

Manual f-test forward selection

This method works by constructing a model, *lmnull*, containing no predictors. Next, the `add1` function is used to determine which variables can be added to the *lmnull* model to assist in predicting the dependent variable. One variable is added at a time, and the output of the model is re-assessed. The output produced is the same as for the backward selection previously explained, containing the degrees of freedom, sum of squares, RSS, AIC, F value

and *p value*. From this output, only significant predictors can be added to the *lmnull* model. Eventually no more variables will be significant, and you have your final model.

```
#build null model
lmnull <- lm(baml ~ 1, data = mds)
#start your variable selection
add1(lmnull, scope = ~temp+rh+hrbk+mthbk+teoml+teoml.l1+teoml.l2+
teoml.l24+nephl.l1+nephl.l24+pm10l+pm10l.l1+pm10l.l2+pm10l.l24+lco+lno2+
lno+ws+wd, test = "F", data=mds)

#start adding significant variables until the output tells you that no
other variables are significant
add1(update(lmnull, ~ +
temp+rh+hrbk+mthbk+teoml+teoml.l1+teoml.l2+teoml.l24+nephl.l1+pm10l.l2+pm1
0l.l1+lco+lno+ws+lno2), data = mds, scope = ~
temp+rh+hrbk+mthbk+teoml+teoml.l1+teoml.l2+teoml.l24+nephl.l1+nephl.l24+
pm10l+pm10l.l1+pm10l.l2+pm10l.l24+lco+lno2+lno+ws+wd, test = "F", data =
mds)
```

The output for the above code is shown below:

```
Single term additions

Model:
baml ~ temp + rh + hrbk + mthbk + teoml + teoml.l1 + teoml.l2 +
    teoml.l24 + nephl.l1 + pm10l.l2 + pm10l.l1 + lco + lno +
    ws + lno2
          Df Sum of Sq    RSS    AIC F value  Pr(>F)
<none>                 3510.9 -23724
nephl.l24  1   0.28946 3510.7 -23724  1.3056 0.25321
pm10l      1   0.62014 3510.3 -23725  2.7975 0.09443 .
pm10l.l24  1   1.13214 3509.8 -23727  5.1078 0.02383 *
wd         1   1.32587 3509.6 -23728  5.9822 0.01446 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For the sake of producing a parsimonious model, we chose to leave out the PM_{10} lag 24 and wind direction, as they were only slightly significant, and would not have improved the model very significantly.

Therefore, the final model using the f-test forward selection is the same as the one made using the f-test backward selection process.

Appendix 7: Output of model with BAM lagged variables included in model.

```
Call:
lm(formula = baml ~ temp + rh + hrbk + mthbk + baml.l1 + baml.l2 +
    baml.l24 + teoml + teoml.l1 + teoml.l24 + nephl.l1 + nephl.l24 +
    pm10l.l1 + pm10l.l2 + lco + ws, data = mds)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4173 -0.0879  0.0235  0.1319  4.7875

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.0246526  0.0659219  15.543  < 2e-16 ***
temp        -0.0046399  0.0008090  -5.735 9.92e-09 ***
rh          -0.0025231  0.0002366 -10.665  < 2e-16 ***
hrbkb        0.0383478  0.0097038   3.952 7.79e-05 ***
hrbkc       -0.0747491  0.0096302  -7.762 8.86e-15 ***
hrbkd        0.0433336  0.0097929   4.425 9.71e-06 ***
mthbkb      -0.0368736  0.0083805  -4.400 1.09e-05 ***
baml.l1      0.6726584  0.0078855  85.303  < 2e-16 ***
baml.l2     -0.1206055  0.0071695 -16.822  < 2e-16 ***
baml.l24     0.0648063  0.0055421  11.693  < 2e-16 ***
teoml        0.0444099  0.0080860   5.492 4.03e-08 ***
teoml.l1     0.0501411  0.0100079   5.010 5.50e-07 ***
teoml.l24    0.0223955  0.0079339   2.823  0.00477 **
nephl.l1     0.1788482  0.0078321  22.835  < 2e-16 ***
nephl.l24   -0.0351865  0.0069020  -5.098 3.47e-07 ***
pm10l.l1    -0.1042262  0.0150639  -6.919 4.72e-12 ***
pm10l.l2     0.1033962  0.0123999   8.338  < 2e-16 ***
lco          0.3074318  0.0297052  10.349  < 2e-16 ***
ws           0.0206971  0.0026440   7.828 5.27e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3432 on 16086 degrees of freedom
  (3546 observations deleted due to missingness)
Multiple R-squared:  0.6695, Adjusted R-squared:  0.6692
F-statistic:  1811 on 18 and 16086 DF,  p-value: < 2.2e-16
```

Above is the output of the model produced that included BAM lagged values. NO, NO_2 and TEOM lag 2 were removed from the hourly model as they were not longer significant. However, NEPH lag 24 became significant, and is included in the model. This model significantly reduces the autocorrelation in the residuals, producing a better fitted model than the hourly model not containing any BAM lagged covariates. BAM lag 1 and BAM lag 2 express an R^2 value of 0.76. There was the risk that including both of these values would produce an overfitting of the model due to multicollinearity, however it was decided that the cut-off for R squared values was 0.8. Therefore, both variables were included in the model.

Appendix 8: Model performance evaluation statistics

In the subsequent definitions, let O_i indicate the ith observed value, and M_i indicate the ith modelled value, for a total of n observations. These definitions are drawn from Carslaw and Ropkins (2012).

***Fraction of predictions within a factor of two observations**, FAC2.*

The fraction of modelled values within a factor of two of the observed values, are the fraction of model predictions that satisfy the following:

$$0.5 \leq \frac{M_i}{O_i} \leq 2.0$$

***Mean bias**, MB.*

The mean bias is a good indicator of the mean over or under estimation of predictions, in the same units as the quantities being considered. It is calculated as follows:

$$MB = \frac{1}{n}\sum_{i=1}^{N} M_i - O_i$$

***Mean gross error**, MGE.*

The mean gross error offers a good tool for the measure of the mean error, regardless of whether it is an over or under estimation. The mean gross error is in the same units as the quantities being considered, and is calculated as follows:

$$MB = \frac{1}{n}\sum_{i=1}^{N} |M_i - O_i|$$

***Normalised mean bias**, NMB.*

The normalised mean bias is a useful tool for comparing pollutants that cover different concentration scales. It is normalised by dividing by the observed concentration. It is calculated as follows:

$$NMB = \frac{\sum_{i=1}^{n} M_i - O_i}{\sum_{i=n}^{n} O_i}$$

***Normalised mean gross error**, NMGE.*

The normalised mean gross error ignored whether the prediction is an over or underestimate, and can be calculated as follows:

$$NMGE = \frac{\sum_{i=1}^{n} |M_i - O_i|}{\sum_{i=n}^{n} O_i}$$

***Root mean squared error**, RMSE.*

This is a statistic typically used providing a good overall measure of how close modelled values are to predicted values, and is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (M_i - O_i)^2}{n}}$$

***Correlation coefficient**, r.*

The (Pearson) correlation coefficient provides an indication of the strength of the linear relationship between two variables. The correlation coefficient is calculated as follows:

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{M_i - \bar{M}}{\sigma_M} \right) \left(\frac{O_i - \bar{O}}{\sigma_O} \right)$$

***Coefficient of efficiency**, COE.*

The coefficient of efficiency is based on Legates and McCabe (1999), (2013), which is simple and easy to interpret. A perfect model has a COE of 1, with no lower bound. A COE of 0.0 indicates the model is no more competent of predicting the observed value than the observed mean can, meaning the model has no predictive advantage. For negative values, the model is less effective than the observed mean in predicting the variation in the observations. The COE is calculated through the following formula:

$$COE = 1.0 - \frac{\sum_{i=1}^{n} |M_i - O_i|}{\sum_{i=1}^{n} |O_i - \bar{O}|}$$

***Index of agreement**, IOA.*

The index of agreement is based on Willmott et al. (2012), ranging between -1 and + 1. Values that nearing + 1 indicate a better model. The IOA can be calculated as follows:

$$IOA = \begin{cases} 1.0 - \dfrac{\sum_{i=1}^{n} |M_i - O_i|}{c \sum_{i=1}^{n} |O_i - \overline{O}|}, when \\ \sum_{i=1}^{n} |M_i - O_i| \leq c \sum_{i=1}^{n} |O_i - \bar{O}| \\ \dfrac{c \sum_{i=1}^{n} |O_i - \bar{O}|}{\sum_{i=1}^{n} |M_i - O_i|} - 1.0, when \\ \sum_{i=1}^{n} |M_i - O_i| > c \sum_{i=1}^{n} |O_i - \overline{O}| \end{cases}$$

Appendix 9: Checking assumptions for daily ARDL model.

Checking symmetry of variables

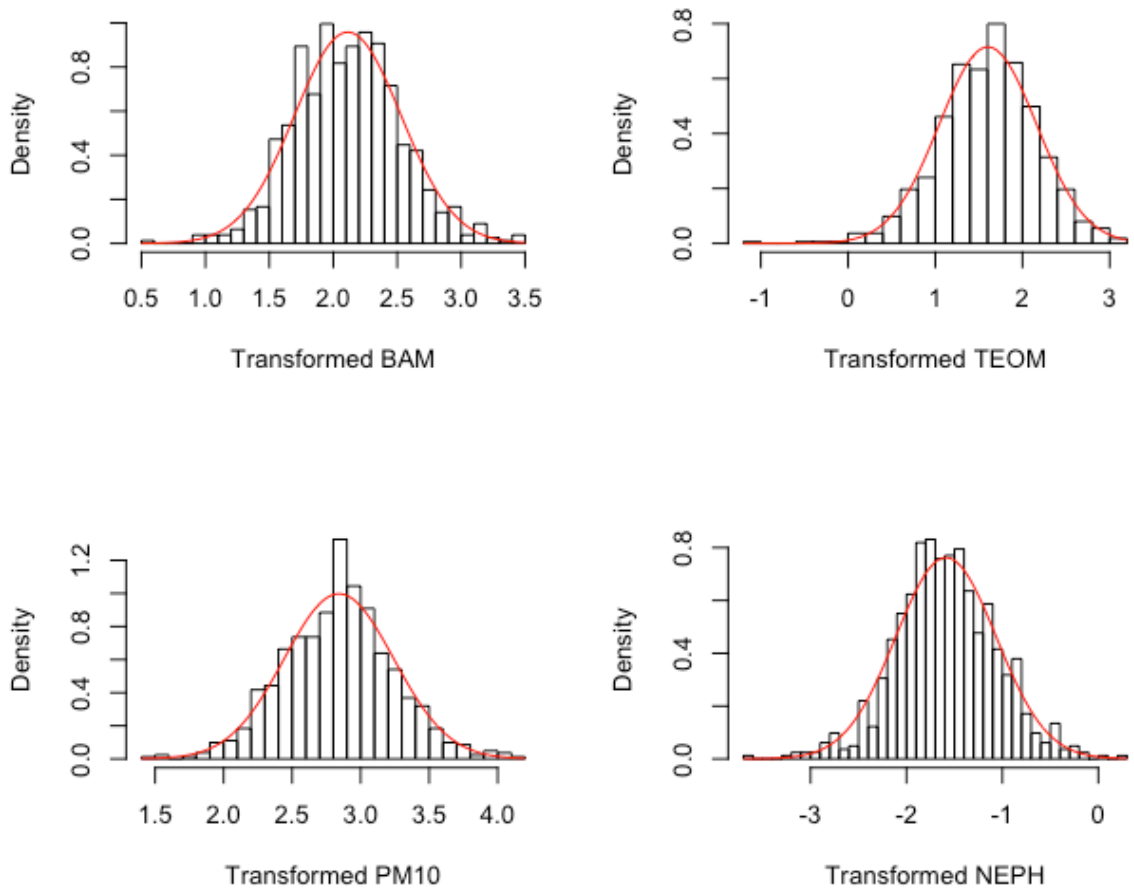


Figure AP9- 1. Density histograms showing symmetry of BAM, TEOM, PM_{10} and NEPH was transformed by a straight log transformation. A normal density curve is fitted to the distribution, using the mean and sample standard deviation to define this particular normal distribution curve.

Checking for linearity of variables

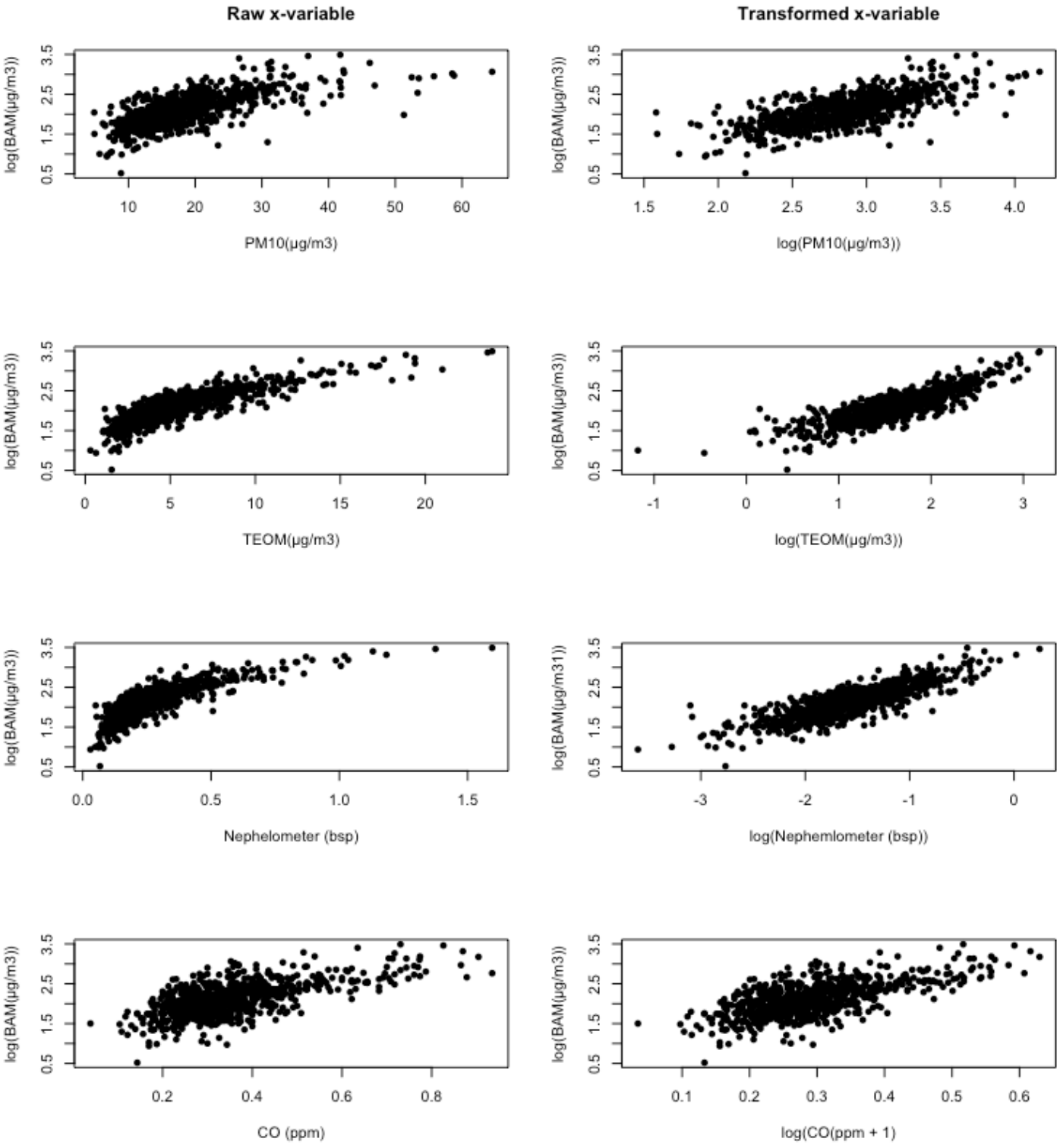


Figure AP9- 2. Checking for linearity of transformed x-variables against transformed BAM. These variables display an improved linearity with the BAM variable once transformed.

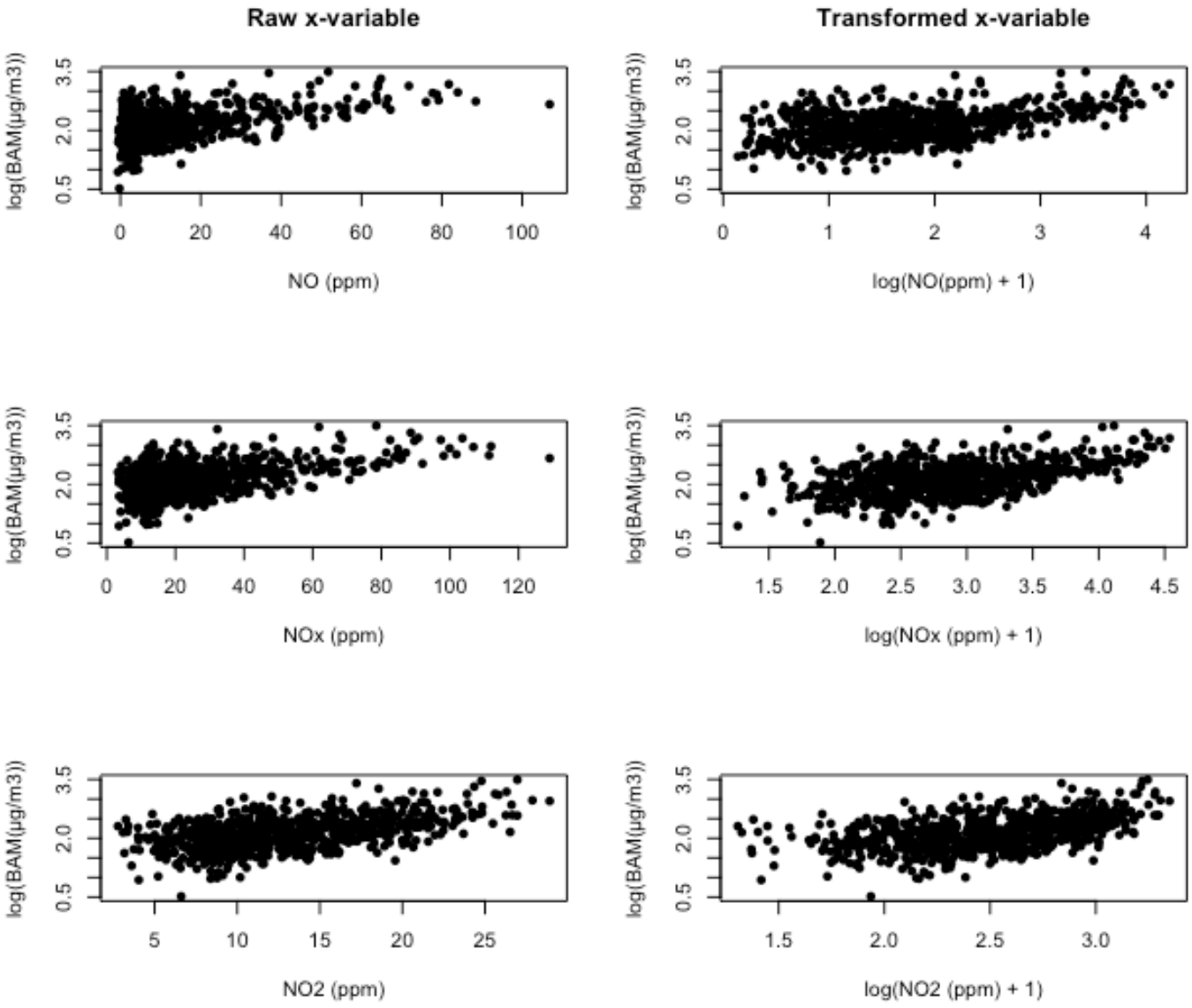


Figure AP9- 3. Checking for linearity of transformed x-variables against transformed BAM. These variables display an improved linearity with the BAM variable once transformed.

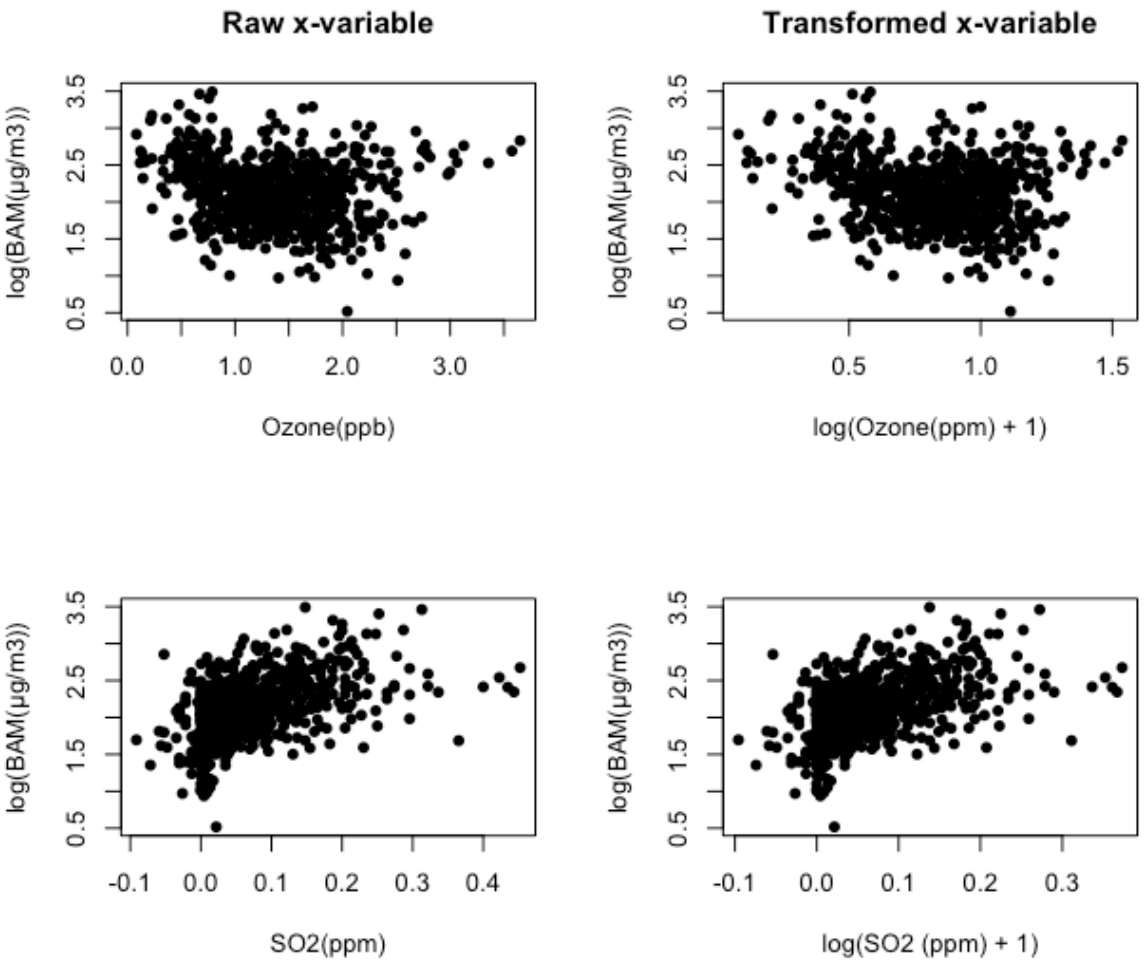


Figure AP9- 4. Checking for linearity of transformed x-variables against transformed BAM. These variables do not display an improved linearity with the BAM variable once transformed.

Testing for multicollinearity of variables

Table AP9- 1. Cross correlation matrix showing the correlation between variables, for daily averaged data. Variables with a correlation of rho ≥ 0.6 are highlighted in yellow, indicating that caution should be used if using both of these parameters in a model as they may possess multicollinearity. Variables with a correlation of rho ≥ 0.8 are highlighted in red. These pairs should not be used in a model together as they will definitely produce overfitting as a result of multicollinearity. The correlation of BAM, TEOM, NEPH, PM_{10} and gasses are calculated from transformed values.

	BAM	TEMP	RH	TEOM	TEOM lag 1	NEPH	NEPH lag 1	PM10	PM10 lag 1	CO	Nox	NO2	NO	Wind Speed
BAM	1.00	0.14	0.06	0.86	0.59	0.84	0.55	0.72	0.45	0.65	0.50	0.47	0.47	-0.34
TEMP	0.14	1.00	0.02	0.20	0.18	0.13	0.12	0.18	0.13	-0.07	-0.38	-0.36	-0.39	0.06
RH	0.06	0.02	1.00	0.00	-0.02	0.27	0.11	-0.27	-0.19	0.43	0.18	0.16	0.16	-0.34
TEOM	0.86	0.20	0.00	1.00	0.56	0.86	0.44	0.78	0.42	0.63	0.49	0.46	0.46	-0.43
TEOM lag 1	0.59	0.18	-0.02	0.56	1.00	0.53	0.86	0.43	0.78	0.32	0.17	0.18	0.15	-0.15
NEPH	0.84	0.13	0.27	0.86	0.53	1.00	0.57	0.65	0.40	0.68	0.50	0.51	0.43	-0.47
NEPH lag 1	0.55	0.12	0.11	0.44	0.86	0.57	1.00	0.33	0.65	0.36	0.18	0.22	0.13	-0.13
PM10	0.72	0.18	-0.27	0.78	0.43	0.65	0.33	1.00	0.50	0.34	0.37	0.35	0.36	-0.12
PM10 lag 1	0.45	0.13	-0.19	0.42	0.78	0.40	0.65	0.50	1.00	0.13	0.10	0.14	0.07	-0.03
CO	0.65	-0.07	0.43	0.63	0.32	0.68	0.36	0.34	0.13	1.00	0.79	0.71	0.77	-0.65
Nox_x	0.50	-0.38	0.18	0.49	0.17	0.50	0.18	0.37	0.10	0.79	1.00	0.94	0.95	-0.61
NO2	0.47	-0.36	0.16	0.46	0.18	0.51	0.22	0.35	0.14	0.71	0.94	1.00	0.80	-0.60
NO	0.47	-0.39	0.16	0.46	0.15	0.43	0.13	0.36	0.07	0.77	0.95	0.80	1.00	-0.55
Wind Speed	-0.34	0.06	-0.34	-0.43	-0.15	-0.47	-0.13	-0.12	-0.03	-0.65	-0.61	-0.60	-0.55	1.00

Appendix 10: Monthly cut-off points for daily data

```
Call:
lm(formula = baml ~ mth + temp + rh + nephl + nephl.l1 + pm10l +
    pm10l.l1 + lco + lno2 + ws + wdir, data = daily)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83864 -0.10915  0.00965  0.11996  0.76613

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.085643   0.227740   9.158  < 2e-16 ***
mth02       -0.042778   0.035359  -1.210   0.2267
mth03       -0.019721   0.035729  -0.552   0.5811
mth04       -0.080914   0.043113  -1.877   0.0609 .
mth05       -0.123916   0.051436  -2.409   0.0162 *
mth06       -0.137599   0.053391  -2.577   0.0102 *
mth07       -0.132969   0.055638  -2.390   0.0171 *
mth08       -0.113153   0.052437  -2.158   0.0313 *
mth09       -0.101748   0.046916  -2.169   0.0304 *
mth10       -0.075524   0.039974  -1.889   0.0592 .
mth11       -0.029686   0.035342  -0.840   0.4012
mth12        0.048130   0.035493   1.356   0.1755
temp        -0.003387   0.003494  -0.969   0.3327
rh          -0.005256   0.001055  -4.980 7.93e-07 ***
nephl        0.388047   0.029510  13.150  < 2e-16 ***
nephl.l1     0.088968   0.021376   4.162 3.53e-05 ***
pm10l        0.236555   0.033345   7.094 3.10e-12 ***
pm10l.l1     0.022070   0.025290   0.873   0.3831
lco          1.682827   0.158746  10.601  < 2e-16 ***
lno2        -0.033865   0.032952  -1.028   0.3044
ws           0.032729   0.013629   2.401   0.0166 *
wdire        0.028117   0.046385   0.606   0.5446
wdirse       0.010026   0.043292   0.232   0.8169
wdirs        0.085242   0.042311   2.015   0.0443 *
wdirsw       0.093174   0.044649   2.087   0.0373 *
wdirw        0.048398   0.049152   0.985   0.3251
wdirnw       0.087015   0.076558   1.137   0.2561
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1813 on 726 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared:  0.8148,  Adjusted R-squared:  0.8082
F-statistic: 122.9 on 26 and 726 DF,  p-value: < 2.2e-16
```

A model was created in R with all possible variables included in the model, with the output of this model shown above. From the output, the cut-off points for the monthly blocks were determined based on significance levels (*p-values*). Months 5 through to 9 were all significant with a *p-value* of < 0.05. These months were blocked as group *b*. All other months were grouped as block *a*, all possessing a *p-value* of > 0.05.

Appendix 11: Two daily predictive models; one using only NEPH and one using only TEOM as the only predictor variables.

An ARDL model was constructed in R, using only NEPH as the covariate used to predict BAM value. The model was applied on the daily data from 03/09/2010 to 29/11/2012.

The measures of predictive ability shown in table Table AP11-1. An adjusted R^2 of 0.71 indicates a good model. However, a model with only the TEOM as the predictor variable produces an adjusted R^2 of 0.74. Therefore, both models are tested to see which has a better predictive ability of actual BAM values. From here on, let the model with only the nephelometer as a covariate be referred to as the 'NEPH only' model, and that with only TEOM as a covariate be referred to as the 'TEOM only' model. A lower AIC, BIC and CV, and higher adjusted R^2 of the TEOM only model indicates a better model than the NEPH only model (Table AP11-1).

The ACF and PACF plots of the residuals indicate there is some autocorrelation in the residuals for both models, as indicated by the lags exceeding the blue dotted 95% confidence line in Figure AP11-1.

We must check that the models are homoscedastic. The residuals occur randomly around the residual line, roughly forming a horizontal band around the zero line, with no one residual standing out from the pattern of the residuals (Figure AP11-2). Therefore, we conclude that both models are homoscedastic.

Prediction and confidence intervals are shown in Figure AP11-3. For the NEPH only model, the 95% confidence interval of the mean predicted transformed BAM is between 2.0868 and 2.1294. The prediction interval for transformed BAM is 1.6670 and 2.549. For the TEOM only model, the 95% confidence interval of the mean predicted BAM is between 2.0821 and 2.1270. The prediction interval is 1.6907 and 2.5135. Prediction and confidence intervals for both models are very similar. The R^2 of the actual BAM vs the modelled BAM for the NEPH only model is 0.71. For the TEOM only model, the R^2 is higher, at 0.75, indicating a better agreement between the fitted and actual values.

Table AP11- 1. Measures of predictive ability of NEPH only model and TEOM only model.

	CV	**AIC**	**BIC**	$\overline{\mathbf{R}}^{\mathbf{2}}$
NEPH only	0.05	-2333.01	-2319.03	0.71
TEOM only	0.04	-2429.62	-2415.65	0.74

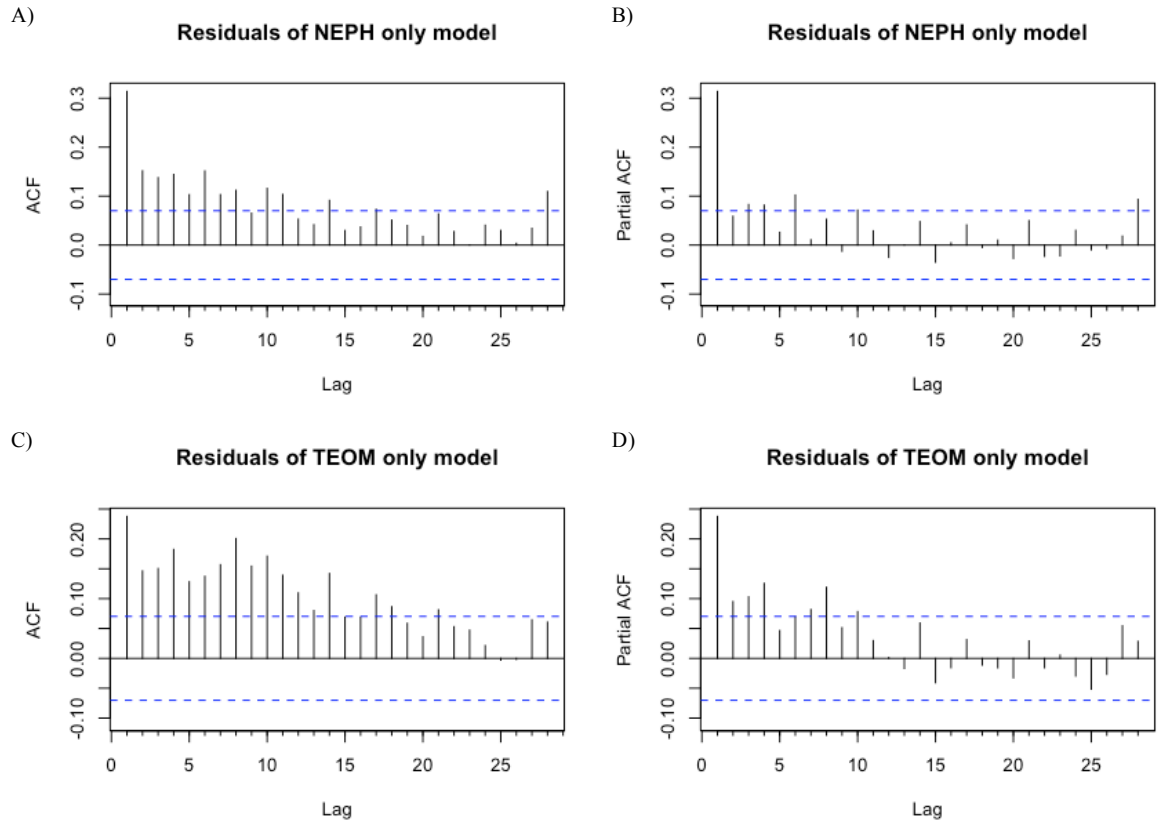


Figure AP11- 1. ACF and PACF plots of residuals of the predictive model using the NEPH only (A and B) and TEOM only (C and D) models, for the prediction of daily BAM values.

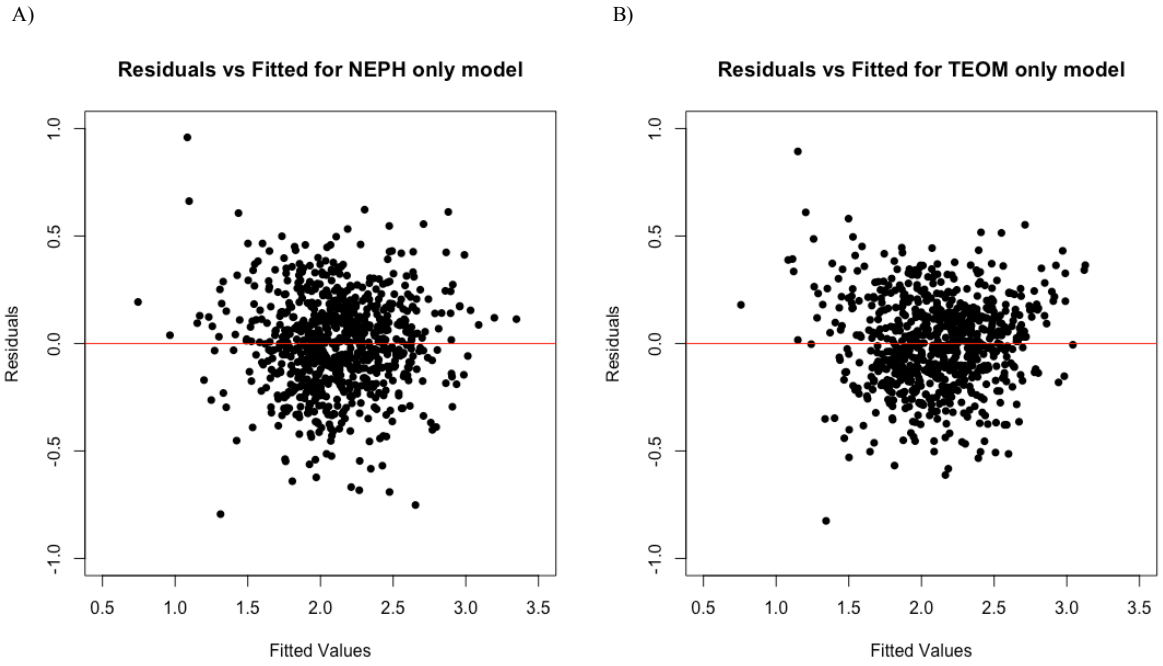


Figure AP11- 2. Plot of the residual vs the fitted values for the daily predictive model for the A) NEPH only and B) TEOM only models. The red line has a slope of 0 along the y-intercept of 0.

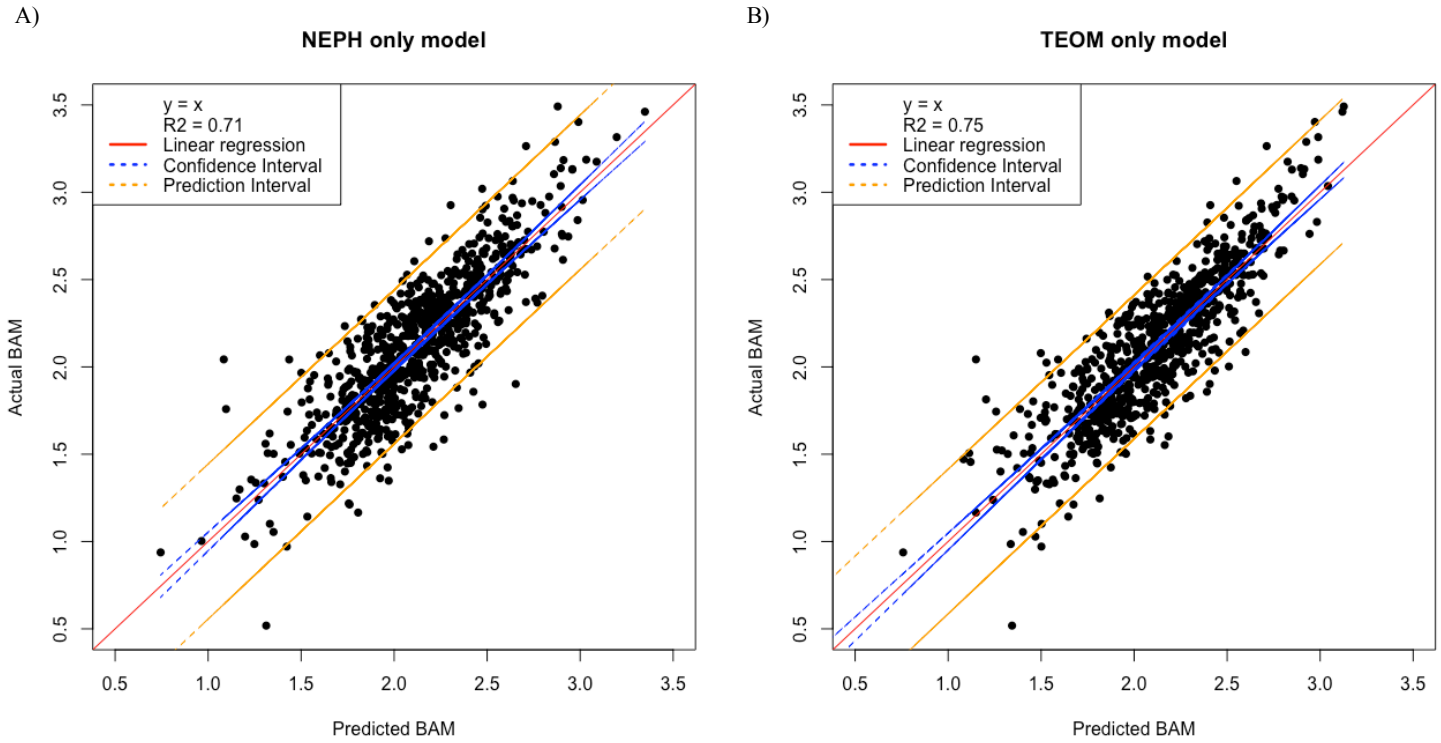


Figure AP11- 3. Linear regression of actual and predicted BAM values over the collocated period for the A) NEPH only and B) TEOM only models. Confidence intervals (blue), prediction intervals (orange), linear regression (red) and R^2 value and coefficients are shown.

One-step time series cross validation was used to evaluate the models performance. The same procedure previously used for the hourly and daily model, and explained in Chapter 4, was used here. The model was developed on daily data from 03/09/2010 to 03/09/2011. The model was then applied, and re-defined each day, up until 29/11/2012.

The `modStats` function from the *Openair* package was used to statistically evaluate the model built on the time-series cross validation. The output for both models is shown in Table AP11-2. Compared to the daily model with no limitations on variable input, both models do not perform as well. The mean bias is greater (NEPH only = 0.150 µg/m^3 and TEOM only = 0.236 µg/m^3, compared to 0.019 µg/m^3), the RMSE is greater (NEPH only = 2.090 and TEOM only = 1.862, compared to 1.642), the COE is lower (NEPH only = 0.510 and TEOM only = 0.541, compared to 0.598), the IOA is lower (NEPH only 0.750 and TEOM only = 0.770, compared to 0.799) and the overall correlation is lower (NEPH only = 0.915 and TEOM only = 0.905, compared to 0.860) (Table AP11-2 compared to Table 6-6).

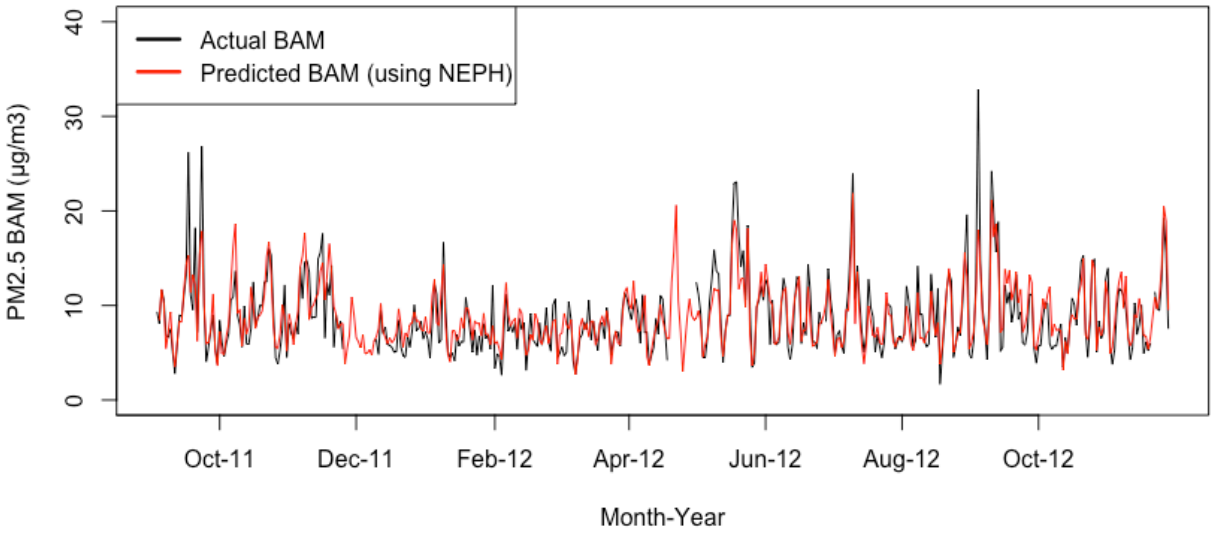
Table AP11- 2. Common numerical model evaluation statistics, based on predicted values from daily one-step time series cross validation, for the NEPH only model and TEOM only model.

	Season	n	FAC2	MB ($\mu g/m^3$)	MGE ($\mu g/m^3$)	NMB	NMGE	RMSE	r	COE	IOA
NEPH only	Autumn	80	1.000	-0.330	1.410	-0.040	0.160	1.910	0.900	0.540	0.770
	Spring	173	0.990	0.390	1.680	0.040	0.180	2.510	0.840	0.500	0.750
	Summer	83	1.000	0.680	1.400	0.100	0.200	1.720	0.720	0.170	0.590
	Winter	91	0.990	-0.370	1.240	-0.040	0.140	1.630	0.900	0.580	0.790
	All data	427	1.000	0.150	1.480	0.020	0.170	2.090	0.860	0.510	0.750
TEOM only	Autumn	80	1.000	-0.168	1.415	-0.019	0.158	1.772	0.925	0.539	0.769
	Spring	172	1.000	0.134	1.418	0.014	0.150	2.043	0.913	0.579	0.790
	Summer	83	1.000	0.974	1.529	0.140	0.220	1.870	0.716	0.098	0.549
	Winter	91	0.989	0.114	1.155	0.013	0.131	1.545	0.924	0.605	0.802
	All data	426	0.998	0.236	1.383	0.027	0.158	1.862	0.905	0.541	0.770

The time series plot of the daily predictions using the time-series cross validation demonstrates that the fitted values from both models fit pretty well with the actual recorded values (Figure AP11-4). Same as for the first daily model, the models fails to capture extreme $PM_{2.5}$ events.

An assessment of error, calculated as actual BAM minus the predicted BAM values, shows that there is still some scatter in the error terms for both models (Figure AP11-5 A and C). For the 427 fitted values using the NEPH only model, 180 values over predicted (42.1%) and 247 under predicted (57.9%) the actual BAM values. For the 426 fitted values for the TEOM only model, 157 values (36.9%) over predicted and 269 values (63.1%) under predicted the actual BAM values. More fitted values under-predicted in both of these models than the daily model with no limits on the input values (53.2% under predicted daily model with no limits on input values). The histogram and frequency suggests a normal Gaussian distribution of error terms for both models (Figure AP11-5 B and D), meeting the assumption of normality for the model.

A)



B)

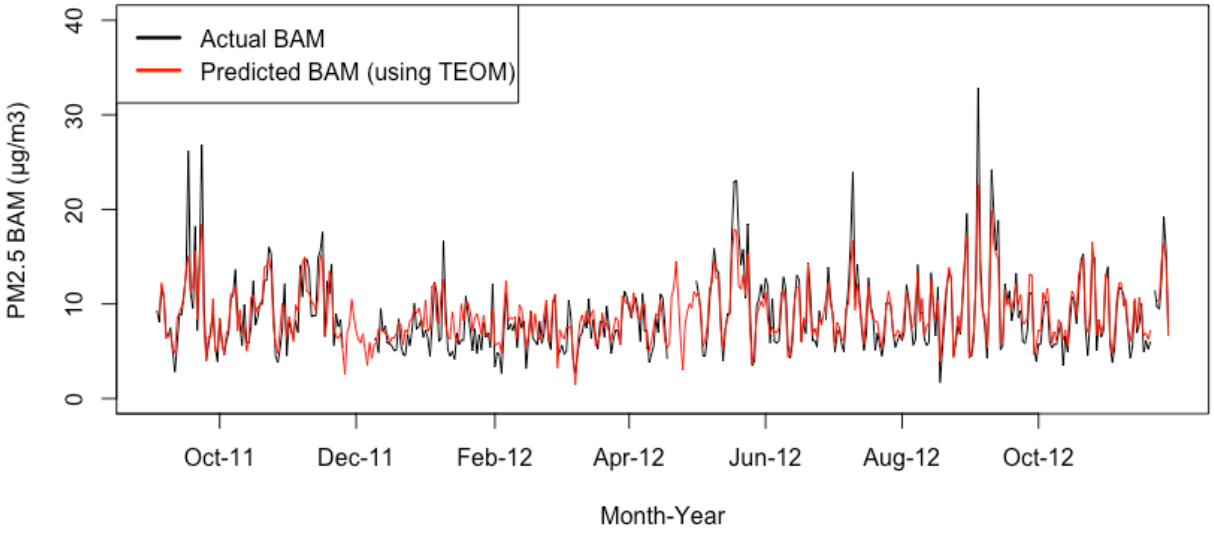


Figure AP11- 4. Time series of actual BAM (black) and predicted BAM (red) values over the period of time when predictions were made using the time-series cross validation, based on daily averages calculated from a predictive model where A) only NEPH and B) only TEOM was used as an independent variable.

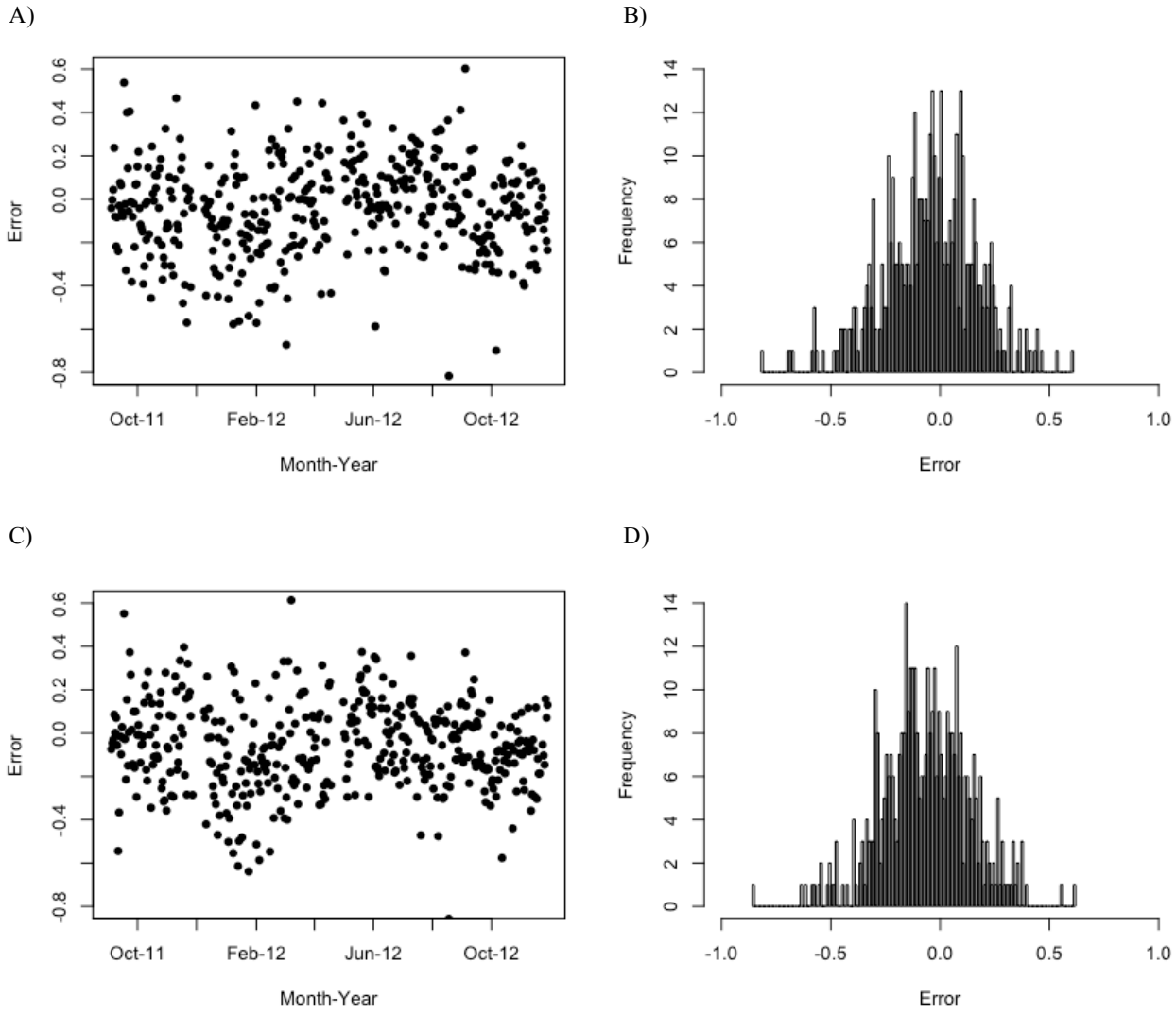
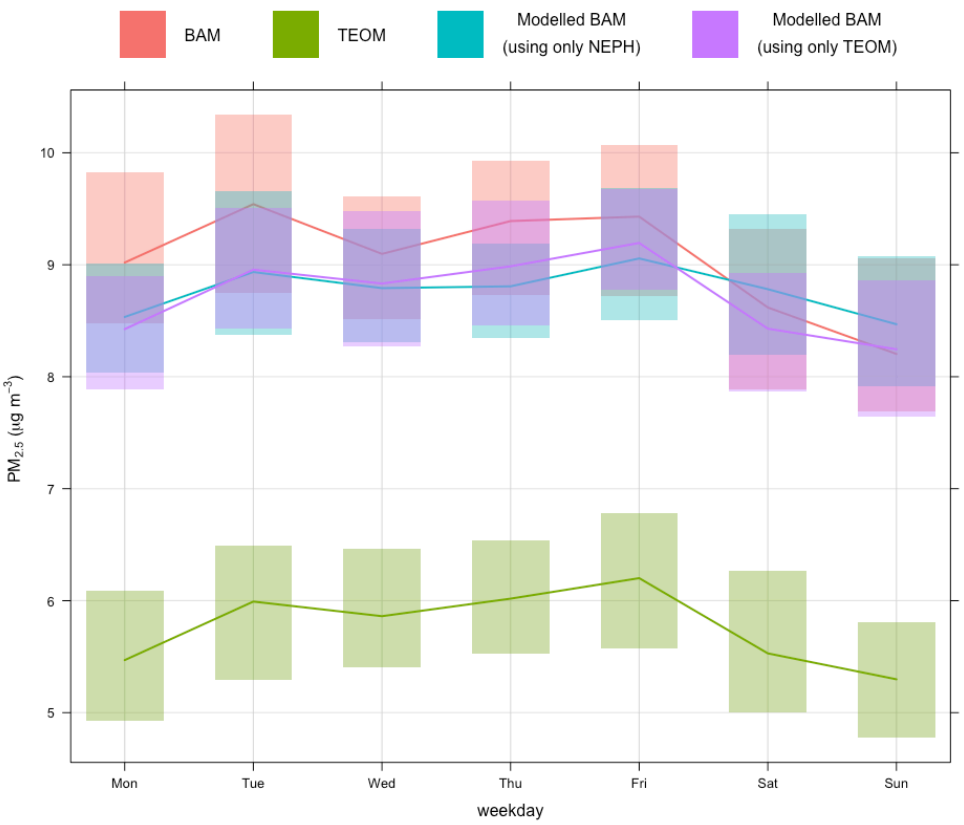


Figure AP11- 5. Distribution of error for daily predictive model using NEPH (A and B) as the only independent variable and TEOM (C and D) as the only independent variable, over the period time where predictions were made using time-series cross-validation. A) and C) depict a time series of the changes in error, and B) and D) show a histogram of distribution of error.

A)



B)

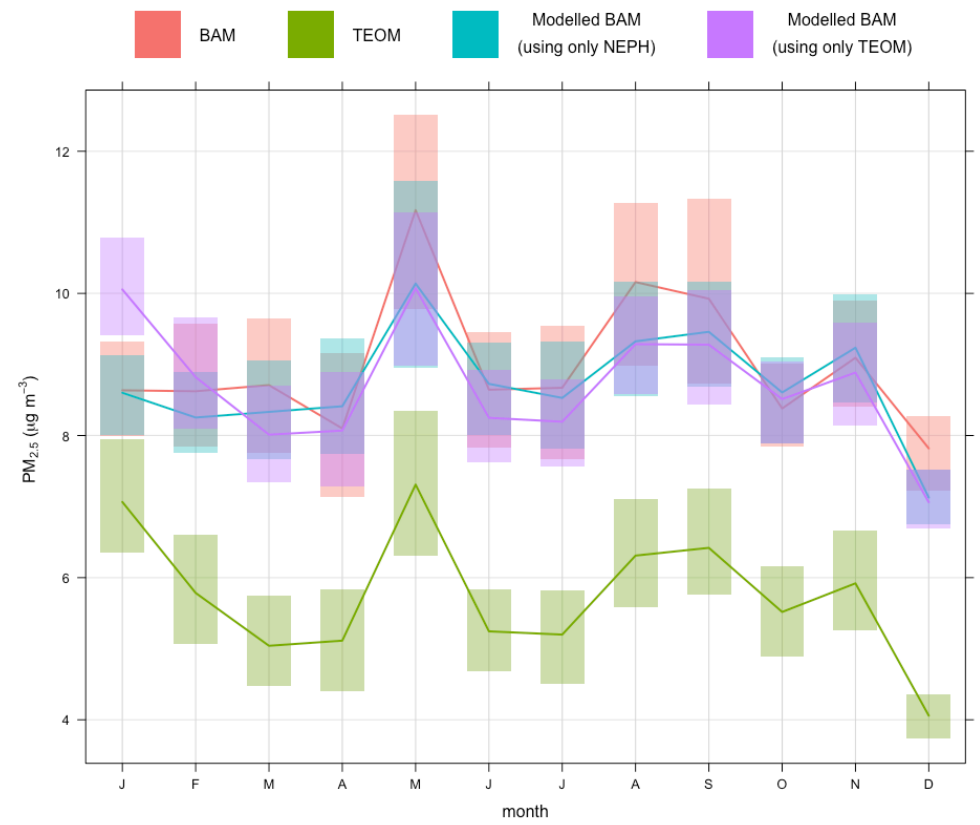


Figure AP11- 6. Time Variation plot showing the actual BAM (red) and TEOM (green) from the collocated period. The modelled BAM values are calculated from a model developed using only NEPH (blue) and only TEOM (purple) as the independent variable. A) shows daily and B) shows monthly average plots. The shading around the boxes indicates a 95% confidence interval.

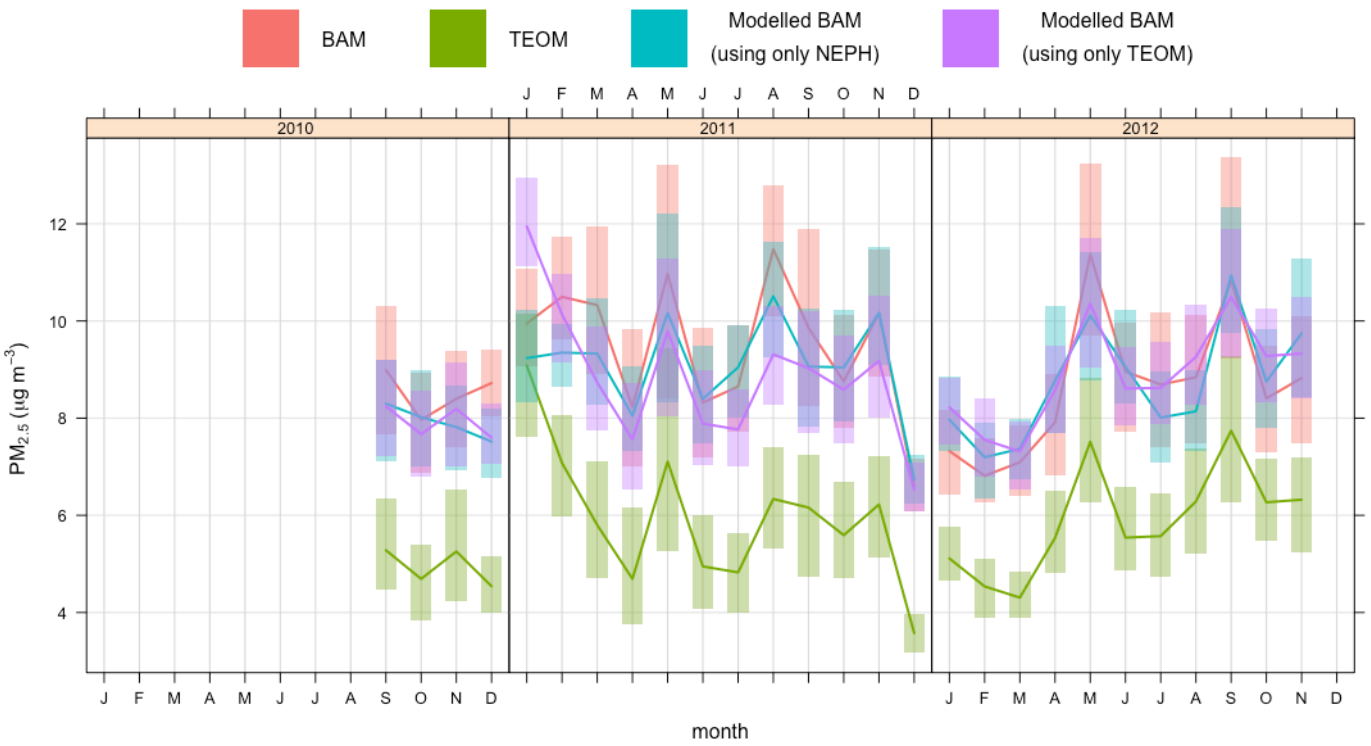


Figure AP11- 7. Time Variation plot showing the actual BAM (red) and TEOM (green) from the collocated period, displayed by year. The modelled BAM values are calculated from a model developed using only NEPH (blue) and only TEOM (purple) as the independent variable. The shading around the boxes indicates a 95% confidence interval.

The predictive ability of both models is ok on a monthly and daily basis (Figure AP11-6). The daily predictive NEPH only model is no better than the TEOM only model. Both of these are not as good as the daily model with no limits on its input variables (see Figure 6-13). Over the collocated period, the model under predicts on a daily basis by approximately 0.5 $\mu g/m^3$, except for Saturday and Sunday where the fitted values are a lot closer to the actual values (Figure AP11-6 A).

On a monthly basis, the models performance is an improvement on what the TEOM was providing (Figure AP11-6 B), but is not as good as the previous daily model with no limit on the input variables (see Figure 6-13). The NEPH only model under predicts for February, March, May, August, September and December, with the remaining months tracking fairly well (Figure AP11-6 B). The TEOM only model over predicts in January, and under predicts in March, May through to December. The modelled BAM values have largely improved from the TEOM values, providing a more true indication of what $PM_{2.5}$ were like over this period, but the previously constructed daily model has a better performance than this model.

When looking at the modelled values for the collocated period per year (Figure AP11-7), it is clear that the TEOM only model fails to capture the January 2011 $PM_{2.5}$ reading,

over-predicting by approximately 2 $\mu g/m^3$. It under predicts the remaining months of 2011, whereas it models a lot closer to the true BAM values for 2012. The NEPH only model appears to be more consistent in its readings.

In conclusion, both of these models have a satisfactory predictive ability, but are not as good as the daily predictive model with no limitations on the covariates used for the model. The summary statistics and time variation plots suggest that the NEPH only model is no better than the TEOM only model. Ultimately, it becomes a trade-off for the user to decide the most appropriate model for their particular application, weighing up the complexity of the model against the predictive ability of the chosen model.